UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

TESIS DOCTORAL

Aplicación de métodos de aprendizaje automático para el estudio de la comorbilidad inversa entre cáncer y trastornos del sistema nervioso central
Application of machine learning methods to study the inverse comorbidity between cancer and central nervous system disorders

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

Kristina Ibáñez Garikano

Directores

Gonzalo Pajares Martinsanz
María Guijarro Mata-García
Alfonso Valencia Herrera

Madrid, 2016

Aplicación de métodos de aprendizaje automático para el estudio de la comorbilidad inversa entre cáncer y trastornos del sistema nervioso central

Memoria que presenta para optar al título de Doctor en Ingeniería en Informática

Kristina Ibáñez Garikano

Dirigida por los Doctores

Gonzalo Pajares Martinsanz
María Guijarro Mata-García
Alfonso Valencia Herrera

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Madrid, 2015

Application of machine learning methods to study the inverse comorbidity between cancer and central nervous system disorders

A thesis submitted in partial fulfillment for the degree of Doctor in Computer Science

Kristina Ibáñez Garikano

Supervised by

Gonzalo Pajares Martinsanz
María Guijarro Mata-García
Alfonso Valencia Herrera

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Madrid, 2015

Itziarrentzat.

’Everything is connected’

Agradecimientos

Cuando queremos acordarnos de un evento del pasado, sólo recordamos lo que rememoramos la última vez. Espero que este hueco me ayude a no olvidar todo lo que ahora recuerdo y sobre todo a la gente que me ha acompañado durante este tiempo.

En primer lugar me gustaría agradecerles a mis directores Alfonso Valencia, Gonzalo Pajares y María Guijarro. A Alfonso por la confianza que tuvo en mi cuando se aventuró al seleccionarme con la beca predoctoral de laCaixa, incluyendo entrevistas telefónicas inter-selváticas. Ha sido el 'revisor' más difícil que he tenido en estos 4 años, he aprendido muchísimo de él, y ha conseguido auto-exigirme en todos los ámbitos. Especialmente, quería agradecerle la paciencia que ha tenido conmigo cuando abandoné el grupo en busca de una motivación aún mayor. Le prometí que acabaría la tesis, y aquí está. A Gonzalo por su implicación, ánimo, trabajo y sobre todo la tranquilidad que me ha transmitido en todo momento. A María por su esfuerzo, sus ideas y sobre todo el ánimo que ha sabido darme en momentos más delicados.

Normalmente cuando uno se embarca en un proyecto de investigación, hay una idea, un cimiento, o datos con los que empezar a hacer algo. No ha sido el caso. Después de varios intentos frustrados de hipótesis biológicas varias que me hice, en el 2011 Alfonso me envió un artículo que acababa de publicar en el Lancet Oncology junto a Rafael Tabarés y Anaïs Baudot entre otros. Se habían leído miles de artículos y habían sacado relaciones en forma de morbilidades entre diferentes enfermedades. Y me dijo: ¿te crees todo esto?. A partir de aquí empecé a exprimir mi cerebro en cada largo que hacía en la piscina, pensando en posibles hipótesis biológicas. Me alié con César, con el que empecé a compartir mis primeras ideas y los datos que iba usando como una primera aproximación. Gracias a César he aprendido muchísimo de estadística, no ha sido fácil trabajar en grupo por los horarios diferentes que llevaba, pero siempre hacía un hueco para contestar serenamente (y con un café) a todas las preguntas banales que tenía. También quiero aprovechar para darle las gracias a Anaïs. Es trabajadora, muy buena compañera, mejor persona y fantástica científica

e investigadora, llena de ideas y nuevos planes con la que he ido creciendo estos últimos años, merci beaucoup! Debo darle las gracias también a Rafa, sin sus trabajos sobre la comorbilidad, su insistencia, ánimos y sobre todo esa energía que da, no podría haber seguido adelante, mila esker! A mis compañeros del CNIO: Jorge, Paolo, Raquel, Jose María, Simone, Juan, Ángel, Edu, Leticia, Txema, Tirso, Dani, David... sobre todo porque hacían que los viernes fueran aún más viernes. En especial a Belén, que consigue contagiar con su simpatía a todos los del laboratorio. Ir a la impresora a por la cosecha es un placer!

Me gustaría incluir en un párrafo a mis compañeros del INGEMM que tanto me han sufrido estos últimos meses. Han sido un apoyo fundamental en este último año. Le quiero agradecer a Ángela no sólo por haberme acogido en el grupo, sino por su apoyo, por los ánimos y sobre todo, por lo que puedo aprender de ella. Además de ser una fantástica bio-informática es una de las mejores jefas que he tenido. A Juan Carlos, Helena, María, Luis Patricia, Ana, Sixto, Fernando, Víctor, gracias por los momentos y el trabajo tan apasionante que compartimos, hacen que ir a trabajar sea un gustazo.

Fuera del ámbito de la ciencia es donde he repuesto todas mis pilas. Agradecerles a mis amigos del sahara las infinitas tardes que he pasado con ellos intentado mejorar el mundo a través de cañas o mostos hepatíticos; Eva (keler killer), Raqui, Fer, Agus, Paloma (Malouma) e Irene mil gracias. Gracias por la paciencia y por la espera Dolo, Sara y Mopita, Blankutxi y Rocio; a las microbiólogas por excelencia Almu y Tere, que me han dado esa energía que sólamente aquellos que han pasado por una tesis doctoral pueden dar, ¡gracias!. Marta (Sobau) y Hodei por haber confiado en mi y sobre todo, por las aventuras que he vivido con ellos y me quedan por vivir. Gracias también a Pol, con quien he descubierto las mejores tortillas de Madrid, los chinos más chinos, haciéndome disfrutar de cada pequeño momento. Clara ha sido uno de los pilares más importantes y fuertes en estos últimos años. Desde que hicimos el máster he compartido apuntes, scripts, artículos, cervezas, toallas, incluso atardeceres y aventeuers con ella. Ha sido mi 'apisonadora' particular yendo siempre dos pasos por delante de mí. Ha sido un apoyo fundamental, tanto a nivel profesional como personal y no tengo palabras para agradecerle todo lo que me ha ayudado. Graciasdankemercieskarrikasko Claruki!

Me gustaría agradecer a mi familia el espacio y libertad que me han dejado y el enorme esfuerzo que ha puesto en mí. A mi aita y ama por haberme entendido en el momento Jablonski-sindromazo y haberme aguantado mis porrusaldas mentales. A Larraitz por las meriendas exóticas y por los viajes en donde 'hacíamos lo que teníamos que haber hecho'. Y a los aitonas por recibirme siempre con una gran sonrisa y alguna que otra tortilla de patata, cada vez que vuelvo a Donosti :).

Grazie mille Alvise, per il tuo sostegno e pazienza. Sei stato il mio postdoc particolare, aiutandomi in ogni momento stressante. Grazie per viaggiare con me, scoprire nuovi posti e anche far niente. Grazie grazie grazie.

Itziar ha sido un apoyo fundamental en mi vida y eso que la perdí a principios de esta etapa. Por ella me enganché a la rutina de la piscina, y tan sólo imaginarme de lo que ella haría en cada situación me da la fuerza necesaria para enfrentarme a lo que sea. No tengo palabras para expresar todo lo que siento por ella. Siempre ha sido mi modelo de persona a seguir y seguirá siéndolo. Acordarme todos los días de ella me da ese empuje invisible que nadie más consigue dar. Era la fuerte y yo la débil y parece que la vida me ha dado un bonus track para poder llegar a ser como ella. Va por ti Itziar. Eskerrik asko.

Abstract

Large-scale biological and medical data is continuously produced, making by now Life Sciences part of Big Data sciences. Such data contain acute information that can enable a better understanding of the molecular mechanisms behind biological systems. This information is essential for the progress in the diagnosis and the treatment of diseases. Bioinformatics and Computational Biology are the disciplines dedicated to the organization, analysis, and interpretation of the data produced by Life Sciences. Indeed, biological data mining is complicated by the heterogeneous and complex nature of the data, that are very dependent on the specific experimental details. The complexity and variety of biological problems require the continuous design, implementation, and application of new methods and algorithms.

This thesis deals with one specific complex biomedical problem, i.e. the lower than expected probability of cancer patients develop some central nervous system (CNS) or neurological disorder and viceversa. This condition is known as *inverse comorbidity*. At the medical level, a better understanding of the connections and interactions between cancer and neurological disorders could potentially improve the quality of life and healthcare outcome of millions of people worldwide. Even if this is a well characterized phenomenon, based on solid population and epidemiological studies, very little is known about the molecular basis of *inverse comorbidity*.

The principal objetive of this thesis is to unveil the biological mechanisms behind cancer-CNS *inverse comorbidity*. We have developed three computational strategies based on machine learning and pattern recognition techniques to address specific problems related with cancer and CNS disorders *inverse comorbidity*. In the first method we advanced a novel data mining approach to compute transcriptomic meta-analyses between some types of cancer and CNS disorders. Using gene expression data, our principal idea was to test whether there was a significant co-occurrence between genes that were up-regulated in cancer and down-regulated in CNS disorders,

and the other way around. The main result of this study was the detection of a significant overlap between the genes up-regulated in CNS disorders and down-regulated in cancers, and viceversa. In addition, a similar outcome was observed at the level of general functions and biological pathways. In other words, this first development set the basis for the study of specific genes and pathways, the up-regulation of which could increase the incidence of CNS disorders and simultaneously reduce the risk of developing cancer, while the down-regulation of another set of genes and pathways could contribute to a decrease in the incidence of CNS disorders while increasing cancer risk.

In the second method we presented a novel computational approach inspired by Simulated Annealing to study the stability of protein interaction networks in cancer and CNS disorders. Molecular systems are organized as networks of interactions. In this case we took advantage of protein–protein interaction networks (PPINs), including information on physical interactions between proteins that form active molecular complexes. In this manner, we integrated gene expression data with PPINs to study the differences in terms of network organization rather than at the level of individual genes. Our proposal was based on the combination of large-scale biological data sets on gene expression data and PPINs. With this information we observed that CNS disorders are characterized by a higher stability while networks informed with cancer gene expression data tend to be less stable than the corresponding controls. Moreover, this instability in the network seemed to increase as cancer evolves.

The third work described a novel methodology inspired by the human reasoning, and based on the combination of supervised and unsupervised classification methods (Self-Organizing Maps and K-Means), and clustering validity indices. Previous studies indicate that cancer related genes tend to encode central hubs within the PPIN, and other authors have shown that cancer related proteins have a much stronger protein–protein interaction density than control proteins in the whole human interactome. With this motivation we examined critically the organization of protein networks around cancer related proteins (CRPs), and compare it with the one of proteins related with neurological disorders (NRPs) pursuing the *inverse comorbidity* theory that relates these complex disorders. Two features were defined, i.e. *dex* and *partial_nE*, that were capable to categorize in different groups proteins related to cancer or CNS disorders, corresponding with clusters with high or low feature values. We observed that relevant clusters enriched in the cancer-related proteins include very connected proteins, while clusters enriched in the proteins related with neurological disorders

encompass less connected proteins. This type of strategy is new in the area of gene expression/protein network analysis applied to the relation between diseases, where it might potentially open the field to new applications.

In conclusion, this thesis constitutes a new and concerted effort to study the molecular basis of *inverse comorbidity*, including three computational methodologies addressing three specific problems in the field. Our results demonstrate that the data mining and machine learning techniques, that have been the drivers in this research, are adequate methodologies to progress in the field of *inverse comorbidity*, as well as the potential of their application to other problems in biomedicine.

Resumen

La cantidad de datos biológicos y médicos que se produce hoy en día es enorme, y se podría decir que el campo de las ciencias de la vida forma parte ya del club del *Big Data*. Estos datos contienen información crucial que pueden ayudar a comprender mejor los mecanismos moleculares en los sistemas biológicos. Este conocimiento es fundamental para el progreso en el diagnóstico y en el tratamiento de las enfermedades. La Bioinformática, junto con la Biología Computacional, son disciplinas que se encargan de organizar, analizar e interpretar los datos procedentes de la Biología Molecular. De hecho, la complejidad y la heterogeneidad de los problemas biológicos requieren de un continuo diseño, implementación y aplicación de nuevos métodos y algoritmos. La minería de datos biológicos es una tarea complicada debido a la naturaleza heterogénea y compleja de dichos datos, siendo éstos muy dependientes de detalles específicos experimentales.

Esta tesis se basa en el estudio de un problema biomédico complejo: la menor probabilidad de desarrollar algunos tipos de cáncer en pacientes con ciertos trastornos del sistema nervioso central (SNC) u otros trastornos neurológicos, y viceversa. Denominamos a esta condición como *comorbilidad inversa*. Desde el punto de vista médico, entender mejor las conexiones e interacciones entre cáncer y trastornos neurológicos podría mejorar la calidad de vida y el efecto de la asistencia médica de millones de personas en todo el mundo. Aunque la *comorbilidad inversa* ha sido estudiada a nivel médico, a través de estudios epidemiológicos, no se ha investigado en profundidad a nivel molecular.

El objetivo principal de esta tesis es explorar los mecanismos biológicos que hay detrás de la *comorbilidad inversa* entre cáncer-enfermedades neurológicas. Hemos desarrollado tres metodologías computacionales basadas en técnicas de aprendizaje automático y reconocimiento de patrones para abordar problemas específicos relacionados con la *comorbilidad inversa*. En el primer método proponemos una nueva

estrategia propia de minería de datos para llevar a cabo un meta-análisis de expresión génica entre ciertos tipos de cánceres y enfermedades neurológicas. A partir de datos de expresión de genes, evaluamos si la *comorbilidad inversa* se relaciona a nivel molecular con una co-ocurrencia significativa de genes que están sobre-expresados en cáncer y sub-expresados en las enfermedades neurológicas, y viceversa. El resultado principal de este estudio confirma la relación entre *comorbilidad inversa* y patrones opuestos de expresión génica, confirmando la existencia de un solapamiento relevante entre los genes sobre-expresados en los trastornos neurológicos y aquellos sub-expresados en cáncer, y a la inversa. Estos patrones génicos inversos se reflejan de modo significativo a nivel de rutas biológicas. Es decir, este primer estudio representa la base para el estudio de genes y rutas biológicas específicas, en el cual los genes sobre-expresados podrían aumentar la incidencia de trastornos neurológicos, y simultáneamente, disminuir el riesgo de desarrollar cáncer. Asimismo, la sub-expresión de algunos conjuntos de genes y rutas biológicas podrían contribuir a la disminución de la incidencia de enfermedades neurológicas, aumentando así el riesgo de cáncer.

En el segundo trabajo presentamos una metodología computacional nueva inspirada en el "Enfriamiento Simulado" (del inglés *Simulated Annealing*) para estudiar la *comorbilidad inversa* entre cáncer y enfermedades del SNC. En este caso utilizamos aproximaciones relacionadas con la teoría de redes, puesto que los sistemas moleculares están organizados como redes de interacción. En este caso, utilizamos las redes de interacción de proteínas (PPINs), que incluyen información de las interacciones físicas entre proteínas que forman complejos moleculares, junto a datos de expresión de genes con los que los integramos para analizar la influencia de la expresión génica propia de cada enfermedad en la organización de la red de interacciones. Observamos que las enfermedades del SNC están caracterizadas por una estructura de red que se puede interpretar de mayor estabilidad (es decir, estado de la red que no se altera significativamente, aun cuando las propiedades fundamentales cambian o se introducen perturbaciones) que sus correspondientes controles, mientras que las redes anotadas con la expresión de genes de cáncer tienden a ser menos estables que sus correspondientes controles. Además, esta inestabilidad en la red parece que aumenta con la progresión del cáncer.

El tercer trabajo describe una metodología nueva inspirada en el razonamiento humano, y se basa en la combinación de técnicas de clasificación no supervisadas (SOM y K-Means). A partir de los principales resultados obtenidos a partir del segundo método, examinamos de manera crítica si las proteínas relacionadas con cáncer

(CRPs) tienden a estar más o menos conectadas. Asimismo, analizamos también si las proteínas relacionadas con trastornos neurológicos (NRPs) tienen un comportamiento opuesto al que muestran las CRPs, siguiendo así la teoría de la *comorbilidad inversa*. Definimos dos atributos, *dex* y *local_nE*, para categorizar y diferenciar en diferentes grupos proteínas relacionadas con cáncer o enfermedades neurológicas, incluyéndose éstas en conjuntos con mayor o menor valor de los atributos definidos. Observamos que los grupos (*clusters*) más relevantes enriquecidos en CRPs incluyen proteínas altamente conectadas, mientras que grupos enriquecidos en NRPs incluyen proteínas menos conectadas entre sí. Este tipo de estrategia es nueva en el área de estudios en la cual se combinan expresión de genes y redes de interacción de proteínas aplicados a la relación entre enfermedades complejas, donde puede abrir el campo a futuras nuevas aplicaciones.

Esta tesis constituye un trabajo de investigación nueva, original y multidisciplinar en el que se estudia la base molecular de la *comorbilidad inversa*, incluyendo tres metodologías que abordan y proponen soluciones a tres problemas específicos en el área. El diseño, desarrollo y aplicación de estas metodologías tienen su eje en los campos de la minería de datos y aprendizaje automático, y los resultados obtenidos son significativos en los campos de la *comorbilidad inversa* y la biomedicina.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Preface

The development of large-scale genomics and post-genomics methods, large biological databases and new statistical approaches in combination with adequate data mining and machine learning techniques allows us to examine for the first time human complex disorder relationships. There is an ongoing debate about the associations between some kinds of cancer and neurological disorders or central nervous systems (CNS) disorders. Certain combinations of cancers and CNS disorders co-occur by chance less often than expected, while others co-occur by chance more often than expected. In medicine the co-occurrence of several diseases is known as comorbidity. This is well stablished at the medical level and is part of the medical protocols. But it is relatively unexplored compared to individual diseases.

The sequencing of the entire human genome allows to identify genes that are causally linked to some diseases. Toward a therapy or cure, whether a particular gene is expressed or not and how the gene functions in normal and affected cells must be understood. The fact is that there is a complex molecular process behind each disease, and it requires an understanding of molecular genetics and the molecular basis of the disease. Each disorder can be considered as having a particular molecular mechanisms, in which some processes underlying normal mechanisms are perturbed and many others are undamaged. The phenomenon of comorbidity often suggests that the molecular mechanism for different disorders intersect. These intersections could be informative about underlying mechanisms and can shed some light on many mechanisms underlying both disorders. In particular, understanding intersections or differences between different diseases might be the key to finding novel treatment for

both types of conditions, for instance thanks to drug repurposing or repositioning.

Epidemiological and clinical population studies can reveal statistical associations between complex disorders. Bioinformatics, data mining and machine learning techniques, and the integrative analysis of massive genomic data, offer an interesting avenue for the understanding of comorbidities while studying, for instance, the participation of genes at the level of biological pathways or networks. Molecular biology can address the molecular mechanisms behind these clinical conditions and might be useful facilitating new treatments and novel drug development, and might help to improve diagnoses and prognosis.

1.2 Evolution of bioinformatics and machine learning

In the advent of high-throughput genomics, datasets need to be managed and interpreted. Bioinformatics and computational biology are the disciplines that encompass the analysis and interpretation of this data, the modeling of biological phenomena, and the development of algorithms (Thampi, 2009). This domain is a multidisciplinary field that includes molecular evolution, biological modeling, biophysics, and systems biology among others.

Despite the fact that bioinformatics seems a recent field of study is in the 1960s when it began and coincided with the rise of molecular evolution. Early contributors to this field include Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle, Richard V. Eck, and Robert S. Ledley. It was in 1962 that Margaret O. Dayhoff wrote *Comprotein*, the first algorithm to determine the primary protein structure (Dayhoff and Ledley, 1962). Few years later, a collaboration between Richard V. Eck, Robert S. Ledley, and others produced the first computer-based collection of protein sequences (Dayhoff et al., 1965) called *The Atlas of Protein Sequence and Structure*. It was the first public comprehensive, computerized, and publicly available database of protein sequences, and it has been used as a model for many molecular databases, such as GenBank[1]. This premier database was an essential tool for the development of molecular biology, molecular evolution, and bioinformatics. Concurrently, advances within the structural

[1]GenBank® is the National Institute Health genetic sequence database, an annotated collection of all publicly available DNA sequences.

biology field were fundamental. First methodologies to analyze and interpret biological data derived from X-ray crystallography to obtain macromolecular structures were developed [for a review see (Cassiday, 2014)]. In an era when computers were not needed to manage data, bioinformatics started bringing a number of fields together in a common pursuit (Doolittle, 2010). Soon, computers started to become more and more important in the handling and analysis of biological data in general.

Figure 1.1: The IBM 7090 computer Margaret O. Dayhoff used in her early work (Hagen, 2000).

A few years later, Dayhoff started analyzing protein evolution using computational methods (Dayhoff, 1969). Meanwhile, Walter M. Fitch was developing algorithms and practical methods for phylogenetic tree construction (Fitch, 1970) and pioneering statistical approaches to sequence comparison (Fitch, 1970) and phylogenetic analysis (Fitch and Margoliash, 1967). He also introduced the covarion approach when studying gene evolution (Fitch, 1976). The principal research interest of Russell F. Doolittle was the evolution of protein structure (Doolittle et al., 1962; Doolittle and Blombaeck, 1964; Doolittle et al., 1967). His work showed how early amino acid sequence comparisons started and revealed a great deal about evolution and how computers started becoming necessary when the number of known sequences began to grow exponentially (Doolittle, 2010). Thus far, publications in this interdisciplinary field have been constant. Moreover, from genomic sequences through proteins, bioinformatics has been applied in various areas, such as molecular medicine, personalized medicine, preventive medicine, gene therapy, drug development, antibiotic resistance, comparative studies, and gene therapy between others.

The first genomes completely sequenced corresponded to bacteria and budding yeast [for a review see (Binnewies et al., 2006)]. The first bacterial genome sequenced was the *Haemophilus influenzae* (Fleischmann et al., 1995), and the budding yeast *Saccharomyces cerevisiae* represented the first eukaryotic genome sequenced (Fraser et al., 1995). These discoveries led to a fast development of new technologies that made easier and cheaper to do sequencing, and the number of complete genome sequences started growing rapidly. It was in 1990 when the Human Genome Project[2] got underway, and since then, a huge amount of biological data has been generated. Thenceforth, the development of tools and methods capable of transforming all this complex data into biological knowledge has been urgently needed (Larrañaga et al., 2006).

Therefore, ways to advance computational methodologies to analyze high throughput data in genomics and proteomics have been extensively studied and are essential in understanding biological mechanisms (Liu et al., 2013a). Thus, machine learning and related techniques such as Markov models, decision trees, supporting vector machines, and neural networks have been increasingly used to solve problems in genomics and systems biology. Markov models have been widely used solving different biological sequence analysis problems such as pairwise and multiple sequence alignments, gene annotation, classification or similarity search (Yoon, 2009). For instance, they have been used in protein secondary structure prediction (Won et al., 2007), gene prediction [for a review see (Wang et al., 2004)], pairwise and multiple sequence alignment [for instance, *ProbCons* multiple sequence aligner which uses hidden markov models to specify the probability distribution over all alignments between a pair of sequences (Do et al., 2005)], homologous protein or nucleotide sequence identification [see the widely used *HMMER* approach (Eddy, 1998), and for a review see (Durbin et al., 1999)], and many others. Supporting vector machines are usually used in the classification and prediction of the biological data since biological databases increase, and automatization of the classification process is needed (Yang, 2004). For example, they have been widely used in protein function prediction [for a review see (Bernardes and Pedreira, 2013)], transcription factor binding prediction (Holloway et al., 2005), pairwise homology [for a review see (Saigo et al., 2011)], or gene expression data classification (Vanitha et al., 2015). Among decision tree algorithms *ID3* (Quinlan, 1986) and its successor *C4.5* (Quinlan, 1993), and *CART* (Breiman et al.,

[2]The Human Genome Project is an international research project to determine the sequence of the human genome and to identify the genes it contains.

1984) are probably the most popular,and they have been applied extensively in computational biology in sequence annotation, as biomarker discovery, regulatory networks, or structural biology among others.The ability of neural networks to learn complex functions from large amounts of data, also makes them an ideal tool to aid in the solving biological problems such as protein three dimensional structure prediction [for a review see (Wu and McLarty, 2000)], sequence feature analysis and classification (Wu, 1997), or coding region recognition and gene identification [for a review see (Rozenberg et al., 2011)].

Interestingly, based on the list of the 100 most highly cited papers of all time (Van Noorden et al., 2014), 7 out of 100 works correspond to bioinformatics, and the *CLUSTAL W* (Thompson et al., 1994), *BLAST* (Altschul, 1997) and, *CLUSTAL X* (Thompson, 1997) machine learning methods appear in the 10th, 12nd, and 28th positions with 40,289, 38,380, and 24,098 cites respectively. In summary, it can be said that machine learning and biology have been feeding each other; and improved computational strategies are necessary to both fields advance.

Machine learning is of huge relevance in bioinformatics and in biomedical science more generally (Jensen and Bateman, 2011). The relationship between biology and the field of machine learning has been long and complex (Tarca et al., 2007). The perceptron algorithm (Rosenblatt, 1958) was the first technique in machine learning that was used to try to model actual neuronal behavior; the field of artificial neural network design grew from this attempt. This algorithm was first used in the molecular biology in the analysis of translation initiation sequences in *Escherichia coli* (Stormo et al., 1982). Unsupervised and supervised machine learning techniques have also been widely used in numerous life science applications. For instance, using gene expression data, patients can be classified in different clinical sets, and new disease groups can also be identified (Perou et al., 1999; Alon et al., 1999; Alizadeh et al., 2000; Ross et al., 2000). Using evolutionary information from multiple sequence alignments to predict protein secondary structure (Rost and Sander, 1994) is another example.

A plethora of biological and medical research problems can be analyzed by integrating advanced machine learning and computational modeling. But rather than focusing on components in isolation (i.e., genes or proteins), the study of the interaction between individual components can help in understanding how biological systems behave. Biological systems can be described as complex networks of biolog-

ically relevant entities (D'Alche-Buc and Wehenkel, 2008). The principal objective of Systems Biology is to understand the network behavior and the dynamic aspects, which requires the use of mathematical modeling. Machine learning is one of the drivers of progress in this context. Systems Biology cares about the study of the systems of biological components (i.e., molecules, cells, organisms or entire species). These systems are dynamic and complex, and their behavior is not trivial to predict from their individual components. Systems Biology consists in the development of mathematical and computational models to describe these dynamical systems. For instance, discovering topological and other characteristics of metabolic networks (Jeong et al., 2000) or analyzing how genetic interactions combining with environmental factors organize development and response to a disease (Bruggeman and Westerhoff, 2007). Also, biological networks have been modeled in multiple works: Geurts et al. (2007) have proposed a method based on kernel trees to predict links in protein–protein interaction networks and enzyme networks.

An interesting point in the evolution of bioinformatics is that in 1951 Alan Turing contributed to mathematical biology with a publication in which he developed the reaction–diffusion theory. This became one of the basic models of theoretical biology and is also considered a foundation of chaos theory (Turing, 1952). This work led to the development of a whole new area of research related to the creation of patterns in nature. Turing discovered, at least in theory, a system based on two molecules that could create patterns of spots or stripes if these molecules were diffused and interacted chemically in a certain way. This theory was accepted as an explanation of simple patterns, such as zebra stripes or the ridges that are formed in sand dunes. Slowly, researchers were piecing together the role of Turing systems in creating biological structures. Interestingly, the Multicellular Systems Biology group at the Centre for Genomic Regulation (Barcelona) coordinated by James Sharpe confirmed that fingers and toes follow the model described by Turing theory (Raspopovic et al., 2014). This study solves the puzzle using systems biology and by combining experimental data and a mathematical model to show which molecules act as Turing predicted, *BMT* or *WNT* genes [see (Sheth et al., 2012; Raspopovic et al., 2014) for more information].

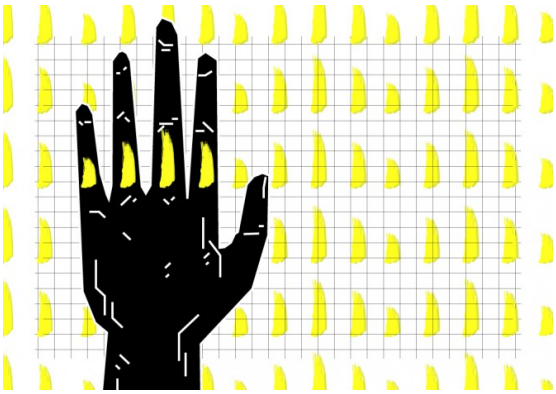


Figure 1.2: Turing system used in creating biological structures (by ©Job Boot).

1.3 Biological concepts

Cells are fundamental building blocks of living organisms and contain different kinds of organelles, such as nucleus, mitochondria, ribosomes, and vacuoles. The nucleus is an essential part, housing the chromosomes that include DNA. DNA is the hereditary material in humans and almost all the organisms that contains the genetic instructions for the development.

In molecular biology, the central dogma is the passage of information from genes (DNA) to proteins via RNA. Figure 1.3 shows an original 1956 depiction by Crick (1958).

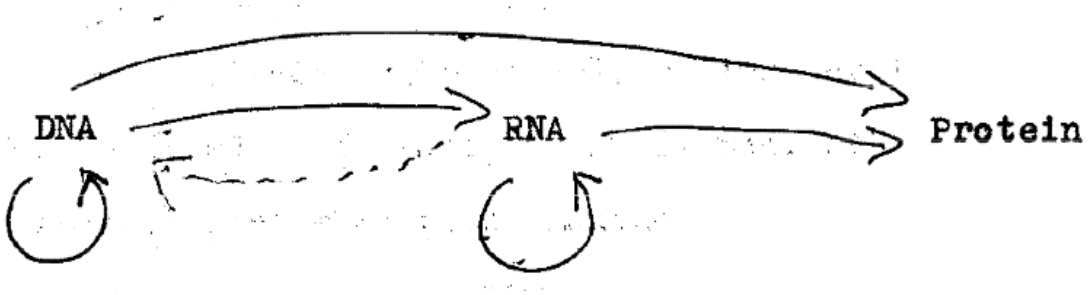


Figure 1.3: An early depiction of the central dogma in molecular biology (Crick, 1958).

The gene has been defined as "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions" (Pearson, 2006; Pennisi, 2007). In other words, genes are the blueprint for life; they tell cells what to do and when to do it. Genes generally express their functional effect through the production of proteins (Figure 1.3), which are essential in order to execute functions in cells. Studying human genes can help us to understand the genetic context behind complex disorders.

1.3.1 Gene expression and microarrays

Transcriptomics or gene expression analysis is the study of the transcriptome, the complete set of RNA transcripts that are produced by the genome in a specific cell. It allows the identification of genes that are differentially expressed in distinct cell populations or in response to different drugs.

Gene expression is a two-step process in which the information encoded in the DNA is used to codify the chain of amino acids that form a protein (Figure 1.3). The control of gene expression causes most phenotypic differences in organisms (Ardlie et al., 2015). Gene expression can explain differences in the phenotype, and so, in the protein function. Because of many disease result from complex changes on the molecular level, observations and models of these processes on the system level are needed.

A measurement of the amount of gene product is sometimes used to infer how active a gene is. An abnormal amount of gene product can be correlated with a deregulation of the transcription that can directly cause anomalous behaviors associated with a disease. For instance, it has been seen that *TOMM34* is frequently over-expressed in colorectal cancer tumors, and is involved in the growth of the colorectal cancer cells (Zhang et al., 2014). Another example are the findings of Barna et al. (2008), which demonstrate that the ability of *MYC* gene to increase protein synthesis, directly augments cell size and is sufficient to accelerate cell cycle progression independently, linking this over-expression of *MYC* gene with an oncogenic signal.

Recent advances in biology have resulted in different techniques and tools to measure gene expression called microarrays. A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. DNA microarrays are used to measure the expression levels of a large number of genes simultaneously or to genotype multiple regions of a genome.

Microarray techniques are used in gene discovery, such as in the global description of genes potentially involved in developmental, physiological, and pathological processes; gene regulation, the description of regulatory networks, based on the assumption that genes regulated in parallel share common control mechanisms; diagnosis, the identification of patterns of gene expression that define disease states and that may represent prognostic indicators, and drug discover, and toxicology.

Gene expression techniques also have several limitations: the output from the analysis of a microarray experiment is usually a large data spreadsheet filled with numbers related to the signal intensity for each gene on the chip. Further analysis is required to identify groups of genes that are similarly regulated across the biological samples under study. It is not a quantitative method, and so it does not perfectly reflect the reality.

1.3.2 Protein–protein interactions

Proteins are large molecules composed of chains of amino acids that form a certain structure, flexible, able to perform structural functions and catalysis. They play crucial roles in living cells and in life; they perform their functions under constant motions and have varied shapes, flexibility, and interactions with other biological molecules. Typically, proteins interact with other proteins, metabolites, and nucleic acids within cells. Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function.

Interactions between proteins are important for the majority of biological functions. Proteins frequently participate in the formation of protein complexes (i.e., group of two or more proteins) and constitute the basis of many biological processes. A protein complex can be considered as a molecular machinery that performs most of the biological functions (Hartwell et al., 1999). For instance, the ribosome, large and complex molecular machine composed by hundreds of proteins, serves as the

site of biological protein synthesis (i.e., translation). Protein–protein interactions are important for virtually every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches, since almost all the drugs are targeted to modify the function of the specific proteins.

These kinds of interactions have been measured using a variety of assays, such as inmuno precipitations and the yeast two-hybrid approach. These techniques have been scaled up to measure interactions on a genome-wide level. High-throughput techniques have also been developed to systematically identify protein complexes using affinity purification techniques followed by mass spectrometry to sequence proteins.

One of the great challenges for molecular biology is to reconstruct the complete network of protein interactions within cells. The analysis of the network biology should also permit scientists to select protein targets for therapeutic intervention by facilitating understanding of the underlying mechanisms of action. Eventually, protein networks may also be used to construct comprehensive dynamic models of molecular interactions within cells, allowing scientists to quantitatively predict the outcome of experiments.

Along with experimental approaches to detect protein interactions, computational methods have also been developed. These methods are used to search for pairs of proteins that have co-evolved, implying that they are likely to be interacting within cells (de Juan et al., 2013; Ochoa et al., 2015). Even though computationally derived interactions are generally not as reliable as experimentally measured ones, they provide a more complete and accurate understanding of protein interactions in combination with experimental data (Mosca et al., 2013).

In summary, proteins do not act alone, and they often form complexes with other proteins to perform a specific task. Existing interactions have been studied, and we use this information to see how those interactions evolve based on the expression of the proteins.

1.4 Meta-analysis

The British statitician Karl Pearson was probably the first to use meta-analytic techniques in order to combine observations from different clinical studies in 1904 (O'Rourke, 2007). In particular, he studied the correlations between inoculations for typhoid fever and mortality for five independent samples (Pearson, 1904). However, the term meta-analysis used to refer to "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" was not coined until 1976 by Gene Glass in his educational research (Glass, 1976). This was followed by articles and textbooks about meta-analysis (Rosenthal, 1978; Cooper and Rosenthal, 1980; Glass et al., 1981; John E Hunter and Jackson, 1982).

The use of meta-analysis in the medical research area began a few years later (Stjernswärd, 1974; Chalmers et al., 1977; Chalmers, 1979) and now its use is widespread. Figure 1.4 shows all the studies in PubMed (pub) that include the word 'meta-analysis' or 'metaanalysis' in their titles between the years 1990 and 2014. Each column represents the number of studies published within each year. The papers published have been counted from a PubMed search with the following commands: meta-analysis[title] OR metaanalysis[title] AND y[dp], y being each year from 1990 to 2014.

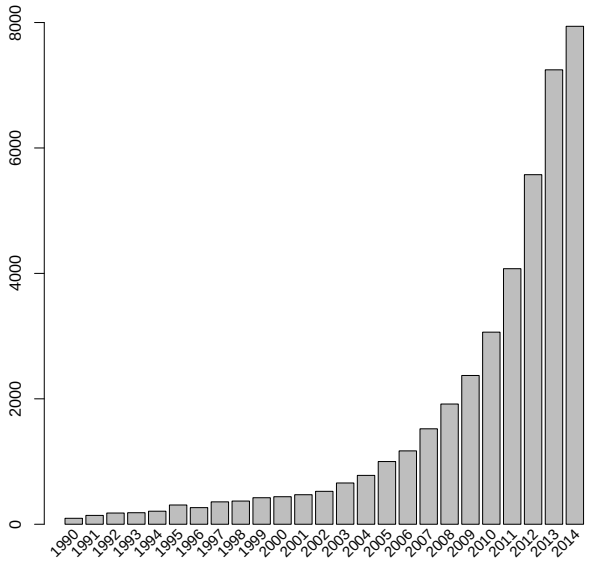


Figure 1.4: Time evolution in the number of publications on meta-analysis.

The principal aim of a synthesis is to understand the results of any study in the context of all the other studies. Meta-analysis is very important for the following reasons:

- **Statistical significance.** The meta-analysis provides a mathematically rigorous mechanism to test the null hypothesis.
- **Clinical importance of the effect.** The meta-analytic approach allows one to compute an estimate of the effect size (e.g., in this work differentially expressed gene (DEG)).
- **Consistency of effects.** Working with a large number of studies is fundamental in determining whether or not the effect size (DEG) is consistent across studies.

1.4.1 Meta-analysis in microarrays

Today, an increasing amount of gene expression data is available in public repositories such as The Cancer Genome Atlas (tcg), GEO from NCBI (geo), and ArrayExpress from EBI (ae). These databases contain valuable information that may lead to new discoveries. From these repositories, one can obtain multiple datasets related to a specific disease. But often, a single study has a small sample size and so its statistical power is limited. Combining information coming from multiple and independent studies linked by the same hypothesis is a practical way to increase the overall sample size and sensitivity, reduce false positives, and provide more robust and validated results. However, this integration is not trivial. The absence of standards for microarray experiments generates heterogeneous datasets, and a direct comparison is not possible (Choi et al., 2003). Even though microarray experiments have been performed in different laboratories with the same research objective, the results of these experiments may differ from each other in many aspects, such as features of the samples, probe sets, or the microarray platform (Shi et al., 2011). Accordingly, the significant genes identified through different experiments may be inconsistent, even using the same statistical analysis.

There are several strategies to handle the combination of such large amounts of data. Several studies directly integrate gene expression data by aligning genes and probes and concatenating samples (Warnat et al., 2005; Xu et al., 2007, 2008), that is, merging the data and then deriving a result (Figure 1.5b). This is, the gene expression values are merged and then, these values are transformed into numerically comparable measures. Meta-analysis is another way of generating more robust and consistent statistical results by integrating multiple datasets (Figure 1.5a). Integrating information from multiple relevant genomic studies has brought new challenges, and microarray meta-analysis in particular has become a frequently used tool in biomedical research in order to increase statistical power (Tseng et al., 2012).

The goal of microarray meta-analysis is to detect DEGs associated with a disease by combining information from several studies. These DEGs might be potential candidate markers for disease classification, diagnosis or prognosis prediction, and help us to understand the genetic mechanisms underlying a disease. There are many different meta-analysis methods that are used to combine information across studies and to generate meta-analyzed p-values.

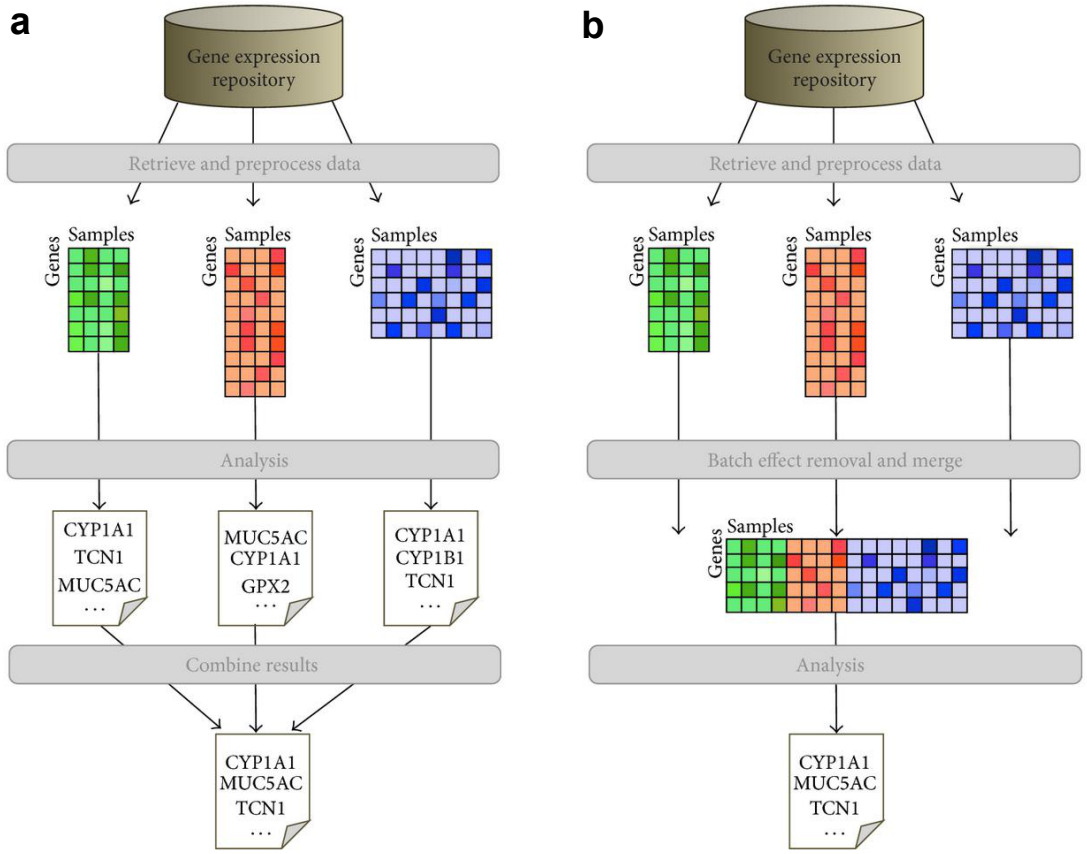


Figure 1.5: Schema of the different approaches in the integrative microarray analysis in the context of the identification of differential expressed genes (DEGs). (a) Meta-analysis. (b) Analysis by data merging [adapted from (Taminau et al., 2014)].

Many microarray meta-analysis methods have been developed and applied in the literature, particularly for DEG detection. Their main objective is to identify DEGs across two or more conditions with statistical significance and/or biological significance. According to a recent review (Tseng et al., 2012), popular methods mainly combine three different types of statistics: p-values, effect sizes, and ranks. The truth is that despite the availability of a large number of methods, the selection of the meta-analysis method depends strongly on the data structure and the hypothesis to achieve the biological goal (Chang et al., 2013). To our knowledge, there are only three comparative studies systematically comparing multiple meta-analysis methodologies (Hong and Breitling, 2008; Campain and Yang, 2010; Chang et al., 2013), and the conclusions are not decisive.

1.5 Comorbidities

Comorbidity or multimorbidity is explained as the presence of additional diseases in relation to an index disease in one individual (Tabarés-Seisdedos et al., 2011). In other words, comorbidity is the presence of several diseases at the same time in one individual.

In the last few decades, epidemiological and clinical studies have been published showing that comorbidity is a universal medical problem because patients with several disorders are the rule rather than the exception (Valderas et al., 2009). The impact of multimorbidity on health is significant. For instance, the research conducted by Fortin et al. (2005) states that in a cohort of more than 700 patients, 9 out of 10 patients had more than one chronic health problem. Along the same lines, an epidemiological study of multimorbidity in Scotland found that almost a quarter of all patients, and more than half of those with a common chronic disorder, have multimorbidity (Barnett et al., 2012). The authors conducted a cross-sectional study in which the medical records of 1,751,841 people from Scotland (more than a third of the population) were used. These records correspond to more than 300 Scottish medical practices and were used to analyze the distribution of multimorbidity (i.e., two disorders from a list of 40 long-term disorders) in relation to age, sex, and socioeconomic status. They found that people with multimorbidity have a reduced quality of life, and that the prevalence of multimorbidity increases with age and socioeconomic deprivation (particularly, physical and mental health comorbidities).

There are two types of comorbidities: direct and inverse comorbidities. Direct comorbidity implies that patients with a particular disease are more likely than expected to suffer other diseases. For example, diabetic patients have a major risk factor for heart disease and hypertension (Bell, 2004), patients infected with hepatitis C have a major risk of liver cancer (El-Zayadi, 2009), and women infected with human papillomavirus are more likely to develop cervical cancer (Hernández-Avila et al., 1997). In contrast, *inverse comorbidity* is characterized by a lower than expected probability of certain diseases occurring in individuals diagnosed with other health conditions (Tabarés-seisdedos and Rubenstein, 2013). For example, studies on cancer risk among patients suffering Parkinson's disease show significantly reduced cancer risk ratios compared to patients without Parkinson's disease (Bajaj et al., 2010).

Many studies analyzing comorbidities between different complex disorders have been published in the last decades. For instance, (Catalá-López et al., 2014a) carried out a meta-analysis of cancer incidence in 577,013 patients from 50 observational studies. They assess the co-occurrence of cancer in patients with CNS disorders, including Alzheimer's disease (AD), amyotrophic lateral sclerosis, autism spectrum disorders, Down's syndrome, Huntington's disease, multiple sclerosis, Parkinson's disease, and schizophrenia. Figure 1.6 sums up graphically the comorbidities they have found between the diseases under study. It shows the relationships among several kinds of cancer and CNS disorders (Parkinson's disease (PD), multiple sclerosis (MS), schizophrenia (SCZ), Down's syndrome (DS), Huntington's disease (HD), and Alzheimer's disease AD)) comorbidities. Red lines represent direct comorbidity cases and green lines represent inverse comorbidity cases.

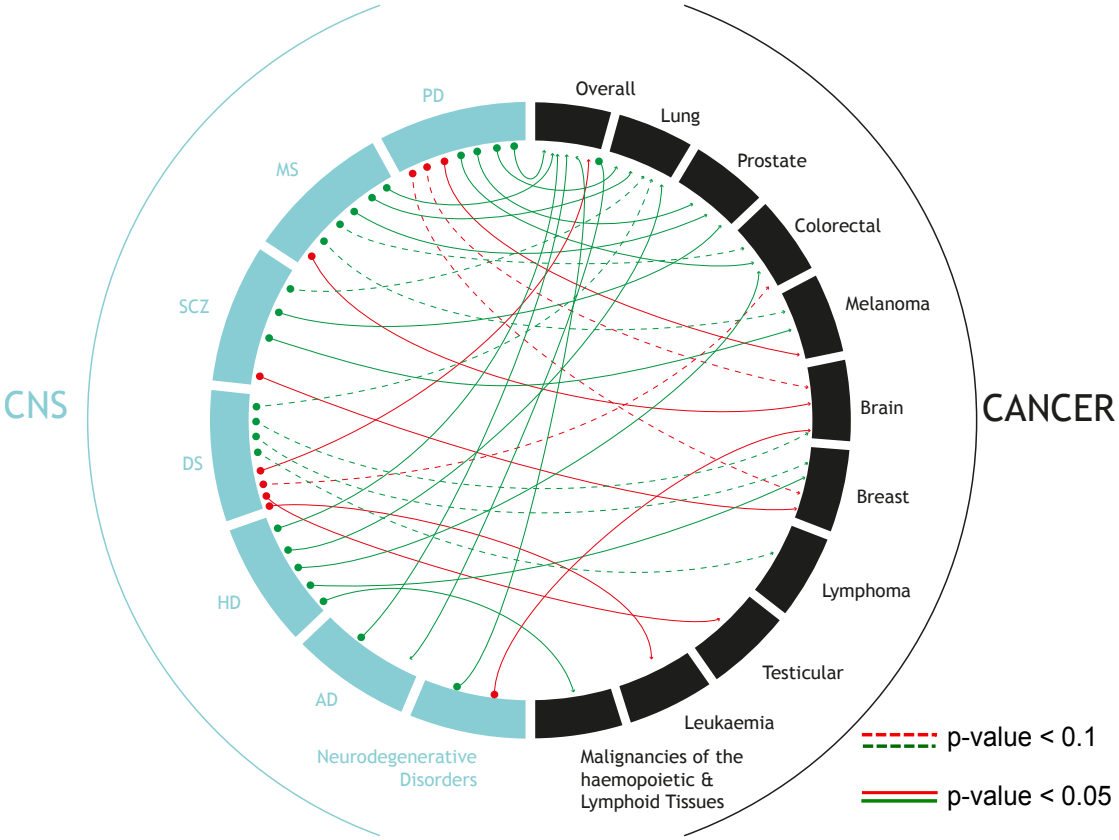


Figure 1.6: Relationships among several kinds of cancer and CNS disorders comorbidities (Figure provided by Dr. Tabarés-Seisdedos). Dotted line represent p-values lower than 0.1, and continuous lines p-values lower than 0.05.

They observe that the presence of any CNS disorder is associated with a reduced co-occurrence of cancer. In particular, patients with Alzheimer's disease, Parkinson's disease, multiple sclerosis, and Huntington's disease have a lower overall co-occurrence of overall cancer (Catalá-López et al., 2014a). By contrast, they observe that patients with Down's syndrome have a higher overall co-occurrence of cancer. Hasle et al. (2000) also found that patients with Down's syndrome have a strong risk of increased co-occurrence of cancer, specifically gastrointestinal and testicular cancers and leukemia.

For both direct and inverse comorbidities, the underlying biological mechanisms are not well described. There are different factors that might account for comorbidity: drug treatments and derived side-effects, clinical factors, an unhealthy lifestyle, the environment, the low incidence of screening, poor access to healthcare, differences in socioeconomic status, and genetic susceptibility (Tabarés-Seisdedos et al., 2011).

1.5.1 Inverse comorbidities

Inverse comorbidity or inverse multimorbidity is characterized by a lower than expected probability of certain diseases occurring in individuals with other health problems (Tabarés-Seisdedos and Rubenstein, 2009). In this work, we focus on this kind of medical condition because it provides a new and potentially better scenario for the study of the biological mechanisms that are behind many complex disorders.

Most of the diseases associated with inverse cancer comorbidity are neurological disorders or central nervous system (CNS) disorders (Tabarés-Seisdedos and Rubenstein, 2009), and they have been previously established by a series of observational studies (Hasle et al., 2000; Catts et al., 2008; Bajaj et al., 2010). This relationship in people with certain CNS disorders is an invaluable opportunity to gain insight into the pathogenesis of these complex diseases, and understanding why certain individuals with CNS disorders are protected against different kinds of cancer could help in developing new treatments, by drug repurposing (Tabarés-seisdedos and Rubenstein, 2013).

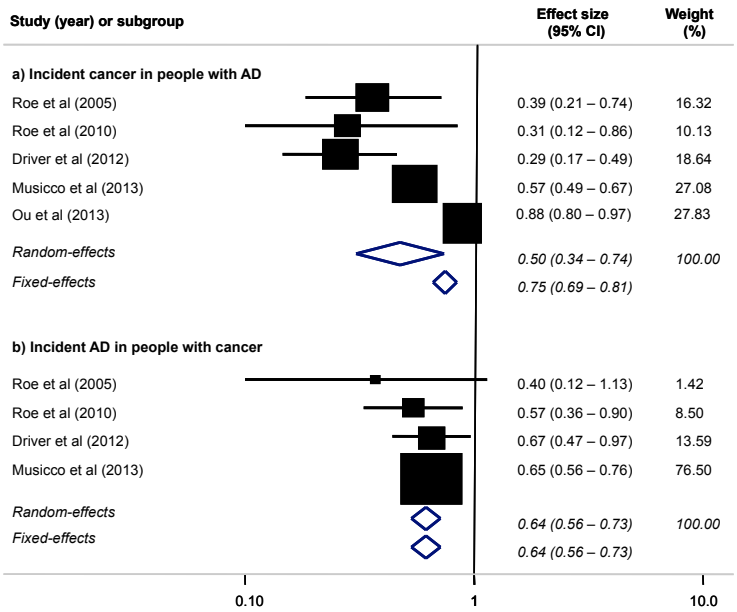


Figure 1.7: Forest plot of (a) the risk of cancer in people with Alzheimer's disease, and (b) the risk of Alzheimer's disease in people with cancer (Catalá-López et al., 2014b) (Figure provided by Dr. Catalá-López).

Several research groups worldwide have published studies related to this condition. Many studies go in the direction of an existing *inverse comorbidity* between cancer and Alzheimer's disease (Behrens et al., 2009; Driver et al., 2012; Thinnes, 2012; Musicco et al., 2013; Ou et al., 2013; Ma et al., 2014; Romero et al., 2014; Driver et al., 2015). For instance, Driver et al. (2012) followed 1,278 participants with and without a history of cancer for 10 years, and 221 cases of Alzheimer's disease were detected. They concluded that patients with a history of cancer have a lower risk of Alzheimer's disease, and patients with Alzheimer's disease have a lower risk of subsequent cancer. A recent study carried out by Catalá-López et al. (2014b) validates and quantifies these epidemiological findings. The authors searched all observational noninterventional studies published up to July 2013 reporting valid measures of the comorbidity-related risk of Alzheimer's disease and cancer. The forest plot (Figure 1.7) shows the risk of cancer in people with Alzheimer's disease and the risk of Alzheimer's disease in people with cancer. These significant findings are in line with those of other studies: there is a mutual protection between Alzheimer's disease and cancer.

Inverse comorbidity between Parkinson's disease and cancer have also been published (Bajaj et al., 2010; Catalá-López et al., 2014a). In particular, Catalá-López et al. (2014a) observed that patients with Parkinson's disease have a significantly reduced co-occurrence of cancer in general and particularly of lung, prostate, and colorectal cancer.

The inverse medical condition found between schizophrenia and cancer is not so clear. Many studies state that a mutual protection between cancer and schizophrenia exists but that it depends on the age (Lin et al., 2013b,a). Cancer risks in patients with schizophrenia are lower for cancers that are more likely to develop at older ages in the general population (e.g., stomach, pancreatic, and prostate cancer), and cancer risks are higher for cancers that have a younger ager of onset (e.g., breast or uterine corpus cancer).

1.6 Motivation

The data generated in Life Sciences are huge and mining such data is twice over. On one hand, data in Life Sciences are heterogeneous. On the other hand, such data are produced massively and repositories are becoming large and large. Data mining and in special, machine learning, is concerned with the development of efficient algorithms and mathematical models to find patterns and statistical dependencies in huge volumes of data. In particular, machine learning methods in Systems Biology are used to explore molecular networks to automatically discover and model properties, interactions, and behavior in biological systems.

Taking advantage of the first available biological data associated with complex disorders is an interesting way to develop different computational approaches to study the molecular bases of the *inverse comorbidity* between different disorders, to analyze the stability of the protein–protein interaction networks in different scenarios, and to classify proteins that might be essential in the disorders under study.

The goal of the thesis is to study the biological mechanisms behind the *inverse comorbidity* between certain kinds of cancer and CNS disorders developing computational approaches based on machine learning and pattern recognition techniques. Understanding why people with certain CNS disorders are protected against some forms of cancer might be the key to finding novel treatments for both types of conditions.

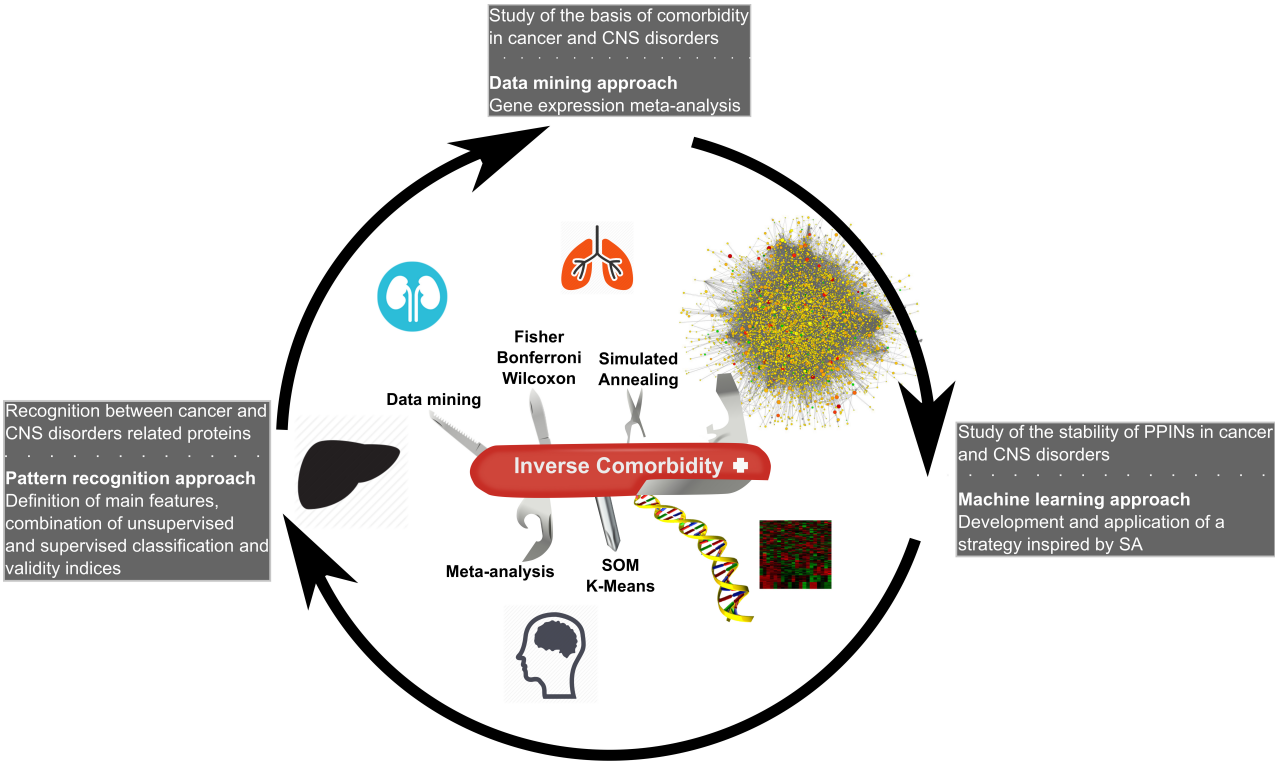


Figure 1.8: Scheme of the thesis.

1.7 Objectives

The main goal is to develop novel computational approaches in order to study diverse perspectives associated with the *inverse comorbidity* between some kinds of cancer and CNS disorders using gene expression data, and combining them with protein–protein interaction network.

The objectives in this thesis are the following:

(a) Study for the first time the molecular processes in common between different types of cancer and CNS disorders using transcriptomic data.

- Implement a computational approach that can be used to conduct transcriptomic meta-analyses between different complex disorders.

(b) Analyze for the first time differences in network stability between cancer and CNS disorders.

- Develop a computational approach inspired on Simulated Annealing to study the stability of protein–protein interaction networks using gene expression data obtained by studying cancer and neurological disorders.

(c) Elaborate a classification strategy to distinguish proteins related to cancer from proteins related to CNS disorders.

1.8 Contributions

1.8.1 Publications in journals related with this thesis

1. Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., and Valencia, A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLoS Genet.*, 10(2):e1004173, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004173

   In this work we present a data mining approach to compute transcriptomic meta-analyses of different complex disorders. In particular, we study three CNS disorders (Alzheimer's disease, Parkinson's disease, and schizophrenia) and three cancer types (lung, prostate, and colorectal) previously described with *inverse comorbidities*. A significant overlap is observed between the genes up-regulated in CNS disorders and down-regulated in cancers, as well as between the genes down-regulated in CNS disorders and up-regulated in cancers. We also observe expression deregulations in opposite directions at the level of pathways. Our analysis points to specific genes and pathways, the up-regulation of which could increase the incidence of CNS disorders and simultaneously lower the risk of developing cancer, while the down-regulation of another set of genes and pathways could contribute to a decrease in the incidence of CNS disorders while increasing the cancer risk. These results suggest that the presenting methodology can be used for other diseases to study comorbidities or other biological hypothesis, and it also might be used with other kind of biological data, different from transcriptomic nature.

   The complete description of this work can be found in Chapter 2.

2. Ibáñez, K., Guijarro, M., Pajares, G., and Valencia, A. A computational approach inspired by simulated annealing to study the stability of protein interaction networks in cancer and neurological disorders. *Data Min. Knowl. Discov.*, 2015. ISSN 1384-5810. doi: 10.1007/s10618-015-0410-5

   In this work we present a novel approach to analyze the stability of protein interaction networks in cancer and CNV disorders. Specifically, the methodology we propose is inspired on the deterministic Simulated Annealing algorithm. Adjusted energy values are used to compare the network stability in disease

and control states in ovarian, colon, liver and kidney cancers, and Alzheimer's disease and schizophrenia. Our results show that cancer networks are less stable than the Alzheimer's disease ones. These results can be interpreted in terms of our previous observations on cancer and Alzheimer's disease *inverse comorbidity*, i.e. Alzheimer's disease patients have lower than expected risk to suffer cancer.

The complete description of this work can be found in Chapter 3.

1.8.2 Communications in conferences

The following works propose diverse computational approaches and data mining based techniques to analyze different aspects regarding the *inverse comorbidity* between some types of cancer and CNS or neurological disorders. Most of them constitute preliminary versions of the works published afterwards or submitted to international journals.

1. Ibáñez K., Guijarro M., Pajares G., Valencia A.
   Stability of Cancer and Alzheimer's Interaction Networks, a Simulated Annealing based Approach
   Conference on Network Biology Community in conjunction with International Society of Computational Biology (ISCB), July 11th 2014, Boston, USA.

   The communication in this conference described a preliminary version of the work described in Chapter 3.

2. Ibáñez K., Boullosa C., Tabarés-Seisdedos R., Baudot A., Valencia A.
   Inverse Comorbidity between Central Nervous System Disorders and Cancers interpreted by Transcriptomic Meta-analyses
   XI Spanish Symposium on Bioinformatics, September 21-24th 2014, Seville, Spain

   The oral communication in this conference presented the final results of the work described in Chapter 2 and in the already published work (Ibáñez et al., 2014).

1.8.3 Other publications

1. Kamieniak, M. M., Rico, D., Milne, R. L., Muñoz Repeto, I., Ibáñez, K., Grillo, M. A., Domingo, S., Borrego, S., Cazorla, A., García-Bueno, J. M., Hernando, S., García-Donas, J., Hernández-Agudo, E., Y Cajal, T. R., Robles-Díaz, L., Márquez-Rodas, I., Cusidó, M., Sáez, R., Lacambra-Calvet, C., Osorio, A., Urioste, M., Cigudosa, J. C., Paz-Ares, L., Palacios, J., Benítez, J., and García, M. J. Deletion at 6q24.2-26 predicts longer survival of high-grade serous epithelial ovarian cancer patients. *Mol. Oncol.*, 2014. ISSN 1878-0261. doi: 10.1016/j.molonc.2014.09.010

   In this work a validation of array CGH is done with external data by means of an adaptation and application of a computational algorithm based on previous strategies. Here predictors of clinical outcome for advanced high-grade serous ovarian carcinoma are studied. The research work consists in analyze 42 ovarian carcinoma samples, evaluating the utility of DNA copy number alterations. In this manner, the loss at 6q24.2-26 is significantly associated with the group of samples of longer survival. This prognostic value is validated in two independent series, one consisting of 36 samples analyzed by fluorescent in situ hybridization and another comprised by 411 ovarian carcinoma samples from the Cancer Genome Atlas (TCGA) repository. This validation with external data corresponds to the computational analysis in the study, a combination of the adaptation and application of the R *DNAcopy* and *CGHcall* algorithms.

2. Pons, T., Paramonov, I., Boullosa, C., Ibáñez, K., Rojas, A. M., and Valencia, A. A common structural scaffold in CTD phosphatases that supports distinct catalytic mechanisms. *Proteins*, 82(1):103–18, 2014. ISSN 1097-0134. doi: 10.1002/prot.24376

3. Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., Rogers, M., Ibanez, K., Boullosa, C., Andres Leon, E., Ben-Hur, A., and Valencia, A. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.*, 41(D1):D142–D151, 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1041

1.8.4 Other communications

1. Ibáñez K., Silla-Castro JC., Martin-Arenas R., Barroso E., Del Pozo A.
   Toward the Characterization of Patterns in Dravet Syndrome combining Next Generation Sequencing, Clinical Data and Machine Learning techniques
   23rd International Conference on Intelligent Systems for Molecular Biology, July 10-14 2015 Dublin, Ireland

2. Silla-Castro JC., Ibáñez K., Lapunzina P., Del Pozo A.
   RAMBO: Really Accessible Management of Bioinformatics Outcome in massive parallel sequencing for clinical studies
   23rd International Conference on Intelligent Systems for Molecular Biology, July 10-14 2015 Dublin, Ireland

3. Del Pozo A., Ibáñez K., Silla-Castro JC., Lapunzina P.
   A non-invasive bioinformatic method for analysis of fetal aneuploidy in maternal blood by Next Generation Sequencing
   23rd International Conference on Intelligent Systems for Molecular Biology, July 10-14 2015 Dublin, Ireland

4. Ibáñez K., Silla-Castro JC., Lapunzina P., Del Pozo A.
   LACONv, Método Para la Detección de Variación en número de copia usando la cobertura de las lecturas de ADN en paneles personalizados de genes en secuenciación masiva
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

5. Del Pozo A., Ibáñez K., Silla-Castro JC., F.Montaño VE., Campos-Barros A., Moreno JC., Heath KE., González-Casado I., Aguado B., Luna A., Nevado J., Vallespín E., Lapunzina P.
   Secuenciación Masiva en la Práctica Clínica: Ajuste de la técnica en una cohorte de pacientes con patología endocrina
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

6. Del Pozo A., Silla-Castro JC., Ibáñez K., Lapunzina P.
   Estudio de los predictores de patogenicidad *in-silico* para guiar el diagnóstico en NGS
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

7. Prieto-Arribas D., Del Pozo A., Ibáñez K., Silla-Castro JC., Vallespín E., Fernández-Montaño VF., Mori AM., Mansilla E., Rodríguez R., Nevado J., de Torres ML., Lapunzina P., García-Santiago F.
   Método no invasivo para el análisis de aneuploidías fetales en sangre materna mediante NGS
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

8. Lucero A.M., Colicheo A., Cambiasso O., Santos-Simarro F., Fernández L., Mena R., Montaño V.F., Ibáñez K., García-Miñaúr S., Nevado J., Silla-Castro JC., Lapunzina P., del Pozo A., Vallespín E.
   Diagnóstico del síndrome de Marfan mediante el estudio del gen *FBN1* por NGS
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

9. Montaño V.F., Fernández L., Ibáñez K., Prior C., Arribas A., Gutiérrez-Larraya F., Oliver-Ruiz JM., García-Guereta L., Deiros L., Martín-Arenas R., Rodríguez-Laguna L., Silla-Castro JC., Lapunzina P., del Pozo A., Vallespín E.
   Panel personalizado de NGS CardioMass_V1.0 en la rutina clínica diagnóstica de cardiopatías
   XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

10. Martín-Arenas R., Fernández L, Ibáñez K., Malpartida SG., Alfonso-Núñez M., International Consortium, Guillén E., Lapunzina P., Barroso E.
    Identificación de nuevas mutaciones en el gen *PCDH19* en pacientes con epilepsia limitada a mujeres con retraso mental asociado (EFMR)
    XXVIII Congreso Nacional de Genética Humana - AEGH, May 13-15 2015 Palma de Mallorca, Spain

11. Ibáñez K., Silla-Castro JC., Lapunzina P., Del Pozo A.
    Detection of Large Copy Number Variation Algorithm using Read-Depth of Coverage in Fitted Panels of Genes
    XI Spanish Symposium on Bioinformatics, September 21-24th 2014, Seville, Spain

12. Del Pozo A., Silla-Castro JC., Ibáñez K., Campos-Barro A., Moreno JC., Heath K., Nevado J., Fdez-Montaño ME., Vallespín E., Lapunzina P.
    Next Generation Sequencing in Clinical Practice: Challenges and Promises in a Cohort of Endocrine Patients
    XI Spanish Symposium on Bioinformatics, September 21-24th 2014, Seville, Spain

1.9 Thesis layout

This thesis is divided in three blocks. Chapter 2 presents a computational approach capable of identifying candidate genes potentially associated with the *inverse comorbidity* through a transcriptomic meta-analysis. The identification of inversely deregulated genes and pathways in complex diseases that have been previously described as inversely comorbid provides, to our knowledge, the first systematic insights into the possible molecular basis of these associations.

Proteins are not islands, and so, they are somehow connected with other proteins. In order to understand the mechanism of the entire system we must analyze the interactions between proteins. In chapter 3 we propose a model that studies the stability of protein–protein interaction networks in cancer and neurological or CNS disorders.

Following the model presented in chapter 3, a pattern recognition is exposed in chapter 4 for protein recognition depending on many structural properties. We propose the definition of features that are capable to identify proteins that are related to some types of cancer and to neurological disorders.

Chapter 5 describes the general conclusions of the work as a whole and sums up the main conclusions of each of the three works that comprise this thesis.

2 Towards the molecular basis of comorbidity between cancer and CNS disorders

2.1 Summary

We present here a novel data mining approach in order to compute transcriptomic meta-analyses in complex disorders, particularly, between some types of cancer and Central Nervous System (CNS) disorders. A significant overlap is observed between the genes upregulated in CNS disorders and downregulated in cancers, as well as between the genes downregulated in CNS disorders and upregulated in cancers. These genes upregulated and downregulated in each disease tissue correspond to a significant major or minor expression between their corresponding healthy tissue, respectively. Moreover, the patients and the tissues are different for cancer and CNS diseases. When these genes are analyzed at the level of pathways, it is observed that they are somehow clustered in pathways with a clear biological significance which goes in the direction of the *inverse comorbidity*. Our analysis points to specific genes and pathways, the up-regulation of which could increase the incidence of CNS disorders and simultaneously lower the risk of developing cancer, while the down-regulation of another set of genes and pathways could contribute to a decrease in the incidence of CNS disorders while increasing the cancer risk. These results have been published in (Ibáñez et al., 2014).

2.2 Introduction

2.2.1 Inverse comorbidity

Epidemiological evidences point to a lower-than-expected probability of developing some types of cancer in certain CNS disorders (Behrens et al., 2009; Devine et al., 2011; Tabarés-Seisdedos et al., 2011; Behrens et al., 2012; Catalá-López et al., 2013; Tabarés-seisdedos and Rubenstein, 2013; Catalá-López et al., 2014b).

Our current understanding of such *inverse comorbidities* suggests that this phenomenon is influenced by environmental factors, drug treatments and other aspects related with disease diagnosis. Genetics can additionally contribute to the *inverse comorbidity* between complex diseases [for review, see (Behrens et al., 2009; Devine et al., 2011; Catalá-López et al., 2013; Tabarés-seisdedos and Rubenstein, 2013)]. This intriguing association represents an invaluable opportunity to understand why certain individuals are protected against different types of cancer and to discover if there are molecular mechanisms that underlie this protection. In particular, we propose here the deregulation in opposite directions of a common set of genes and pathways as a molecular mechanism directly related with *inverse comorbidity*. Using transcriptional data, we tested whether there is a significant overlap between genes that are up-regulated in cancer and down-regulated in CNS disorders and the other way around (Figure 2.1).

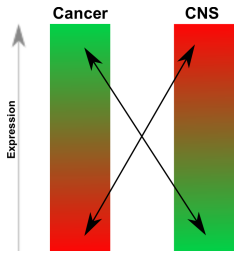


Figure 2.1: Idea.

In particular, based on the study of (Tabarés-Seisdedos et al., 2011), we have focused our study on schizophrenia (SCZ), Alzheimer's (AD) and Parkinson's diseases (PD), and colorectal (CRC), lung (LC) and prostate (PC) cancers (Table 2.1), where the medical observations of *inverse comorbidity* are particularly strong.

	Increased co-occurrence of cancer	Decreased co-occurrence of cancer
Parkinson's disease	Malignant melanoma, and skin, breast and thyroid cancers	Smoking-related cancers (especially lung, colorectal and bladder cancer), and non-smoking-related cancers (especially prostate cancer)
Schizophrenia	Breast cancer	Lung, colorectal and prostate cancers
Alzheimer's disease	None known	All types of cancer
Down's syndrome	Acute leukaemias, testicular cancer, some gastrointestinal cancers, extragonadal germ-cell tumours, and retinoblastoma	Solid tumours (eg, neuroblastoma, medulloblastoma and other embryonal tumours of childhood, and breast and skin cancer in adults)
Diabetes	Pancreatic, liver, colorectal and bladder cancers, and non-Hodgkin lymphoma (type 2 diabetes)	Lung cancer, Hodgkin's lymphoma (type 1 diabetes), and prostate cancer (type 1 and 2 diabetes)
Anorexia nervosa	Non-Hodgkin lymphoma	Breast cancer
Allergy-related diseases	Oesophageal and colorectal cancer, myeloma and prostate cancer	Bladder cancer, Hodgkin's and non-Hodgkin lymphoma, and breast and pancreatic cancers
Multiple sclerosis	Brain tumours	Lung cancer

Table 2.1: Summary of published findings about increased and decreased co-occurence of cancer in people with other complex diseases (Tabarés-Seisdedos et al., 2011).

2.2.2 Meta-analysis for DEG detection

We have performed integrative meta-analyses of collections of gene expression data, publically available for AD, PD and SCZ in CNS disorders, and LC, CRC and PC in cancer. Our objective is to detect differentially expressed genes (DEGs) in expression microarrays of samples labeled with two conditions (e.g., cases versus controls), in a collection of studies with different number of samples. Microarray meta-analysis is usually applied combining multiple studies of related conditions to

better detect these DEGs, increasing the statistical power, reducing false positives and, providing more robust and validated conclusions. For this problem, two major types of statistical procedures have been more frequently used in the literature: combining effect sizes and combining p-values (Song and Tseng, 2014). From a theoretical point of view, there is no single method which performs uniformly better than the others in all datasets for various biological objectives (Littell and Folks, 1971). From a practical point of view, several studies have been published comparing and statistically characterizing different meta-analysis methods with no deciding outcome (Hong and Breitling, 2008; Campain and Yang, 2010; Chang et al., 2013).

In this part of the work, we used the combining effect sizes strategy. The Fixed Effects Model (FEM) and the Random Effects Model (REM) are the most popular methods (Cooper et al., 2009). This approach has many advantages to be applied to microarray data: 1) It provides a standardized index (the measure of gene expression levels is not interchangeable). 2) It can be easily combine different results by integrating microarray data. 3) It is superior to other meta-analytic approaches, due to the power of handle the variability between studies (Choi et al., 2003). FEM and REM are usually more powerful to directly synthesize information of the effect size estimates, compared to p-value combination methods (Song and Tseng, 2014). Indeed, they are particularly applicable to samples with two conditions (cases versus controls) when the effect sizes can be defined and thus, combined.

Accordingly, we have made use of FEM and REM meta-analysis statistical procedures for DEGs detection, identifying genes differentially expressed with statistical and biological significance.

2.3 Material

We propose a strategy to test whether gene expression is involved in the presence of *inverse comorbidity* between different kinds of cancer and CNS disorders. For such aim, we have developed a pipeline which discovers relationships between gene expression data from different kinds of human disorders have been retrieved.

2.3.1 Gene expression data

Gene expression raw data (CEL files) have been downloaded from NCBI GEO omnibus (geo), EBI ArrayExpress (ae) and Stanley Medical Research Institute Online Genomics Database (smr) for CRC, LC and PC in cancers, AD, PD and SZC in neurological disorders, and for asthma, HIV, malaria, dystrophy, sarcoidosis in control diseases. Details of the selected gene expression studies are listed in Table 2.2.

For each disease, experimental gene expression studies have been filtered to select only the ones profiling at least 9 samples for disease and control cases (statistical reasons), with Affymetrix arrays (GeneChip® Human Genome U133 Plus 2.0, GeneChip® Human Genome U133A and GeneChip® Human Genome U133A 2.0 containing 23,945, 14,538 and 14,538 genes, respectively). For CNS disorders, only studies that measure gene expression in brain tissues have been selected. For cancers, only gene expression studies carried out in the LC, CRC and PC tumor tissues have been considered (i.e., blood is not considered). Even though brain tissue is not often available, it has been seen that gene expression in blood is poorly preserved comparing to tissue (Cai et al., 2010).

Table 2.2: Gene expression datasets.

Tissue	Platform	Sample Size	Source
Alzheimer's disease			
Entorhinal Cortex	HG-U133Plus2	23	GSE5281
Hippocampus	HG-U133Plus2	23	GSE5281
Medial Temp. Gyrus	HG-U133Plus2	28	GSE5281
Posterior Singulate	HG-U133Plus2	22	GSE5281
Primary Visual Cortex	HG-U133Plus2	31	GSE5281
Superior Frontal Gyrus	HG-U133Plus2	34	GSE5281
Hippocampus	HG-U133Plus2	24	GSE1297
Parkinson's disease			
Postmortem substantia nigra	HG-U133Plus2	25	GSE7621
Postmortem medial, lateral and frontal	HG-U133A	45	GSE8397
Substantia nigra	HG-U133A	29	GSE20292

Table 2.2 – Gene Expression datasets

Tissue	Platform	Sample Size	Source
Schizophrenia			
Postmortem cerebellum	HG-U133Plus2	28	GSE4036
Postmortem frontalBA46 cortice	HG-U133Plus2	55	Dobrin
Postmortem hippocampus (CA1)	HG-U133Plus2	41	Laeng
Postmortem thalamus (MD)	HG-U133Plus2	26	Kemether
Postmortem frontalBA46 cortice	HG-U133A	67	AltarA
Postmortem frontalBA46 cortice	HG-U133A	50	AltarC
Postmortem frontalBA46 cortice	HG-U133A	67	Bahn
Postmortem frontalBA46 cortice	HG-U133A	69	Kato
Colorectal cancer			
Colon macro- and micro-dissected	HG-U133Plus2	105	GSE20916
Colon tissue	HG-U133Plus2	64	GSE8671
Colon tissue	HG-U133Plus2	81	GSE9348 GSE4183
Colon tissue	HG-U133Plus2	33	GSE7307 (GSM175905) GSE2109
Lung cancer			
Lung tissue	HG-U133Plus2	178	GSE3526 GSE19188
Lung tissue	HG-U133A	54	GSE7670
Lung tissue	HG-U133A	107	GSE10072
Prostate cancer			
Prostate tissue	HG-U133Plus2	122	GSE17951
Prostate tissue	HG-U133A	57	E-TABM-26
Prostate tissue	HG-U133A2	89	GSE6956
Malaria			
Peripheral blood mononuclear cells	HG-U133A	36	GSE5418

Table 2.2 – Gene Expression datasets

Tissue	Platform	Sample Size	Source
Muscular Dystrophy			
Quadriceps muscle	HG-U133A	37	GSE6011
Quadriceps and biceps muscle	HG-U133A	20	GSE11681
Pulmonar Sarcoidosis			
Lung tissue	HG-U133Plus2	12	GSE16538
Asthma			
Circulating CD4+ and CD8+ T-cells	HG-U133Plus2	40	GSE31773
Alveolar macrophage samples from bronchoalveolar lavages	HG-U133A	10	GSE22528
HIV			
Lymph node tissue	HG-U133Plus2	34	GSE16363
Centrum semiovale(deep white matter)	HG-U133Plus2	25	GSE28160

2.4 Methods

We have developed a methodology that follows the steps represented in Figure 2.2. The first stage corresponds to select the most suitable available studies and to retrieve the corresponding gene expression raw data with the filters previously mentioned at section 2.3 (Step 1). Then, gene expression data is preprocessed (Step 2), and a meta-analysis is done (Step 3). As result, for each disease, a consensus gene set of DEGs is obtained. From this point two different aspects are done. Concurrently, a gene-based analysis is done using the Fisher's exact test to compute whether DEGs of a particular disease are enriched in DEGs from other different diseases (Step 4), and a pathway-based analysis is also done (Step 5).

Figure 2.2: Workflow with the steps of the proposed methodology.

A literate programming and dynamic report with a thorough description of our analysis workflow is accessible [see the supplementary material (Ibáñez et al., 2014)]. The completing code contains a step-by-step description of the analysis, and the code can also be executed directly into R. It hence provides the detailed explanations of the tools and the parameters, and allows the complete reproduction of our results.

2.4.1 Microarray data selection (Step 1)

The dataset selection is fundamental. The data quality is one of the most important obstacle in order to achieve a successful meta-analysis (Eysenck, 1994). The use of a poor quality data set in the combination of a large number of studies can greatly attenuate the information contained (e.g., more noise), losing the statistical power or even putting out of true the final biological conclusions (Kang et al., 2012). Up to this point, the amount of studies which combine relevant and homogeneous data in genomic meta-analysis is increasing, but the data quality control is ignored and the inclusion/exclusion criteria usually depends on experts' opinion or on a naïve threshold by sample size or microarray platform.

Kang et al. (2012) propose a methodology called MetaQC, with objective quality control (QC) and inclusion/exclusion criteria for genomic meta-analysis. They propose several quantitative quality control measures, particularly covering accuracy and consistency of DEGs detection. The robustness and effectiveness of the methodology is demonstrated [see (Kang et al., 2012)], and it has been already used in other studies

(Wang et al., 2012; Rung and Brazma, 2013; Manczinger and Kemény, 2013; Chang et al., 2013; Chikina and Sealfon, 2014; Song and Tseng, 2014; Kim et al., 2014).

After having selected gene expression datasets containing at least 9 samples for disease and control cases for statistical reasons with one of the two Affymetrix platforms (see Section 2.3.1), MetaQC R package (Kang et al., 2012) has been used in order to assess the inclusion/exclusion criteria, taking advantage of preprocessed gene expression data. The fact of include an additional informative study to the meta-analysis can provide increased statistical power to detect more DEGs, but adding a lower quality score can deteriorate the performance.

2.4.2 Microarray gene expression normalization (Step 2)

The preprocessing or normalization step is essential to avoid using bad data, distinguish noise and the actual biological data, and to be able to compare data from multiple arrays. The collected microarray data from the different experimental studies are normalized. The raw data consists in an image file with fluorescence intensity values stored. These intensities correspond to levels of hybridization produced by the microarray platform. The preprocessing or normalization means to convert the raw data into useful biological data. Thus, image data should be translate into intensity values, and bias should be also removed.

Accordingly, the raw data are normalized with frozen Robust Multiarray Analysis (fRMA) (McCall et al., 2012) from the R Affy package (Gautier et al., 2004), which removes certain batch effects by down-weighting probes that have high between-batch residual variance [see supplementary material (Ibáñez et al., 2014) for the code].

2.4.3 Microarray gene expression meta-analyses (Step 3)

As it has been previously described in Section 1.4, the meta-analytic techniques are useful to combine observations from large amount of experimental studies for the purpose of integrating the findings. Even though microarray experiments have been performed in different laboratories with the same research objective, the results of these experiments may differ from each other in many aspects, such as features of the samples, probe sets, or the microarray platform (Shi et al., 2011). Accordingly,

a meta-analytic strategy is essential to identify consistent significant genes through different experiments.

There are several strategies to handle the combination of large amounts of data (such as represented in Table 2.2). Particularly, the goal of microarray meta-analysis is to detect DEGs associated with a disease by combining data from several studies. Rhodes et al. (2002) presented an inter-study validation of microarray meta-analysis and the most developed meta-analytic method is combining effect sizes. According Cohen's (1988) definition, the effect size is the magnitude, or size, of an effect. In this case the change of gene expression in each disease is expressed as 'effect size', a standardized index measuring the magnitude of a treatment or covariate effect. An effect size based approach has beneficial features to be applied to microarray data: It provides a standardized index, and it is superior to other meta-analytic methodologies because it has the ability to manage the variability between all the gene expression datasets that are combined.

Hence, we propose a meta-analysis strategy based on the effect sizes because it allows the identification of consistent genes up- and down-regulated in all the experimental studies. The effect sizes are combined to obtain the estimate of the overall mean. Accordingly, the combination of effect sizes gives information about the magnitude and direction (up- or down-) of the gene expression. The principal idea is to compute DEGs for each study, and then work with them to assess the consistency of the effect across studies, computing then a summary effect. This summary effect is the mean of the effect sizes, assigning more weight to the more precise studies (Borenstein et al., 2011). In others words, a DEG is differentially expressed if it is differentially expressed in all the studies, generating a study-invariant list, removing experimental or technical inconsistencies between the different studies (Li and Tseng, 2011).

In order to assess the differential expression of a gene, a standardized mean difference is used as an effect size index (Equation 2.1) (Hedges and Olkin, 1985).

$$d = \frac{\bar{X}_d - \bar{X}_n}{S_p} \tag{2.1}$$

where $\bar{X}_d$ and $\bar{X}_n$ represent the means of the disease (d) and normal (n) groups respectively, and S_p indicates an estimate of the pooled standard deviation. Datasets

with n samples (both cases and controls), the unbiased estimate is obtained as $d' = d - 3d/(4(n-2)-1)$, indicating the correction for sample size bias. The estimated variance of the unbiased effect size is obtained as $\sigma_d^2 = (n_d^{-1} + n_n^{-1}) + d^2(2(n_d + n_n))^{-1}$, where n_d and n_n are the sample sizes of each group (disease, normal) and d is the unbiased effect size (Hedges and Olkin, 1985). This indicates the precision of the measure each dataset or study provides.

Let μ be the overall mean of the expression of a gene in all the datasets (for instance, in the AD), and Y_i the observed effect size for each independent dataset. The general model is given as:

$$\begin{aligned} Y_i &= \theta_i + \epsilon_i, \qquad \epsilon_i \sim N(0, s_i^2) \\ \theta_i &= \mu + \delta_i, \qquad \delta_i \sim N(0, \tau^2) \end{aligned} \tag{2.2}$$

where τ^2 represents the between-study variance (i.e., the variability between the studies) and s^2 describes the within-study variance (i.e., the sampling error conditioned on the ith study). Here Y_i and s_i^2 are given by d and σ_d^2 described previously, and μ is the average measure of the differential expression of a gene across all the datasets.

There are two different approaches when effect sizes are computed:

- **Fixed Effect Model (FEM)**
  FEM model assumes that all the studies in the meta-analysis share a true effect size (fixed effect) and that all the differences in observed effects are due to sampling error (Borenstein et al., 2011), because each study uses a different set of participants. In other words, the factors that could vary the effect size are the same in all the studies within the meta-analysis and so, the effect size is just the same in all the studies. It also follows that the observed effect size might vary among studies due to the random error inherent in each study. In this manner, $\tau^2 = 0$ and consequently, $Y_i \sim N(\mu, s_i^2)$ (Choi et al., 2003).

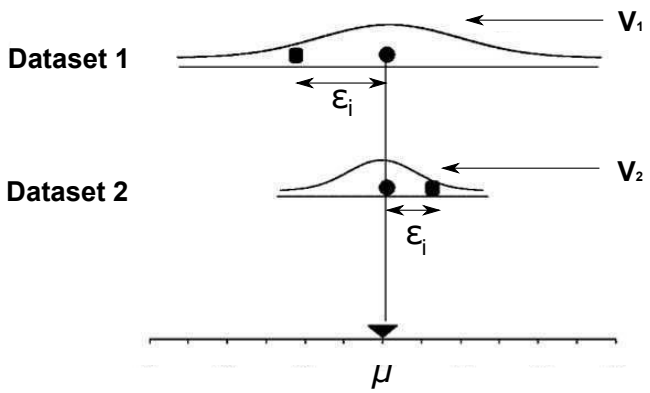


Figure 2.3: Schematic of the FEM model [adapted from (Borenstein et al., 2010)].

The FEM model is depicted in Figure 2.3 with an example of a disease with two gene expression datasets (V_1 and V_2) that share a common (true) effect size (μ). For each study, the true score is represented by a filled circle. The circle for each study falls at the common μ, since all the studies are assumed to share the same effect size. The filled square in each study represents the observed mean, which differs from the true mean because an estimation error. The observed effect size for study i is:

$$Y_i = \mu + \epsilon_i \tag{2.3}$$

where ϵ_i represents the difference between the common mean and the observed mean for the study i. In this way, following the definition of FEM, there is only one way of variation: the estimation error (ϵ_i).

- **Random Effect Model (REM)**
  REM model by contrast, assumes that the true effect could vary among the studies. For instance, the effect size might be lower or higher in studies in which patients are older than the others. It is very common that studies within the meta-analysis have been implemented or developed in different conditions or they differ in the collection of patients, among other reasons. Thus, in REM it is assumed that there might be different effect sizes underlying different studies (Borenstein et al., 2011). In this manner, the effect size is computed from a

distribution with a study-specific mean θ_i and variance s_i^2. Indeed, each θ_i is assumed to be draw from some superpopulation with the overall mean μ and variance τ^2. Accordingly, $Y_i \sim N(\theta_i, s_i^2)$ and $\theta_i \sim N(\mu, \tau^2)$ (Choi et al., 2003).

In this manner, large studies are capable to describe more precise estimations than smaller studies, measuring the overall mean taking into account all the effect sizes. This way avoids that larger studies dominate on the smaller studies.

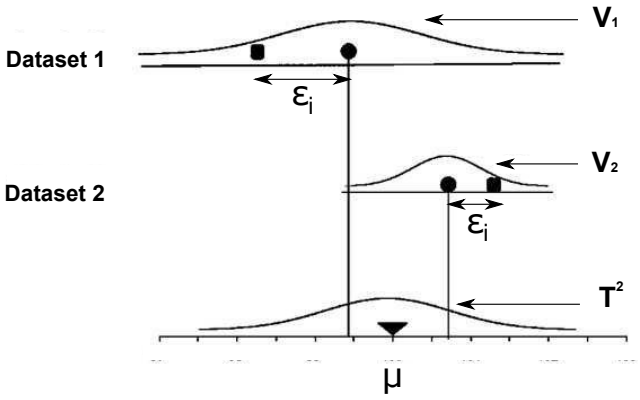


Figure 2.4: REM model
Schematic of the REM model [adapted from (Borenstein et al., 2010)].

The REM model is depicted in Figure 2.4 with an example of a disease with two gene expression datasets or studies (V_1 and V_2). For each gene, the overall mean of the effect size under study is represented by μ, a measure of the average differential expression for that gene. The observed effect size for study i is given by the following equation:

$$Y_i = \mu + \epsilon_i + \phi_i \tag{2.4}$$

where ϕ_i is the difference between the distribution mean (μ) and the true mean (θ_i) for each study i: $\phi_i = \theta_i - \mu$. And ϵ_i is the difference between the true mean for each study i (θ_i) and the observed mean (Y_i, Equation 2.4) for the study i: $\epsilon_i = Y_i - \theta_i$.

On one hand, FEM model considers as true, only sampling error as the reason of variation among effect sizes in the studies under the meta-analysis. Hence, this model is plausible when the studies are close replications of one another (i.e., using similar procedures, measures, etc.). On the other hand, in the REM model is assumed that the studies will differ not only because they have different samples or participants, but also because of divergences in the way they are conducted. Thus, this model is plausible when variation among the studies is evident.

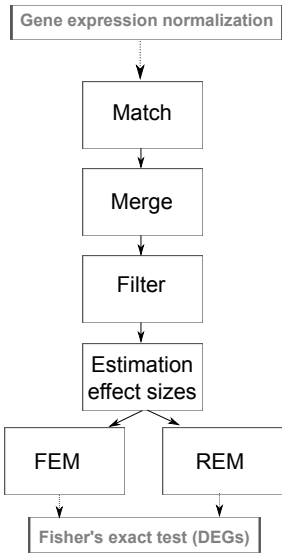


Figure 2.5: The meta-analysis workflow following the R MetaDE package (Wang et al., 2012).

We have followed the pipeline presented in Figure 2.5. Before doing the meta-analysis for each disease, the data must be preprocessed to assess the effect sizes and variances. In this manner, since multiple probes within each microarray platform match to the same gene, probes in each study need to be matched to official gene symbols. When multiple probes match to an identical gene symbol, the probe that presents the greatest inter-quartile range (IQR) is selected to represent the target gene symbol. Large IQR represents greater variability, and thus, greater information content in the data. This is done by using $MetaDE.match$ function from the R MetaDE package (Wang et al., 2012). Then, we have filtered out genes not present in all the studies, and genes with very low gene expression that are identified with small average expression values across the majority of studies by using the $MetaDE.merge$ and $MetaDE.filter$ functions from the R MetaDE package (Wang et al., 2012) respec-

tively. Finally, the effect sizes and the corresponding sampling variances have been calculated for every gene of each study of each disease, following the formulas and the strategy presented in Choi et al. (2003), and they have been combined with FEM and REM models by using the $ind.cal.ES$ and $MetaDE.ES$ functions from the R MetaDE package (Wang et al., 2012) respectively.

Accordingly, using the gene expression normalized data (Section 2.4.2), microarray meta-analyses have been undertaken for each disease independently using the MetaDE R package (Wang et al., 2012), using both models, FEM and REM. MetaDE implements meta-analysis methods for differential expression analysis, and we have used the FEM model (Choi et al., 2003). Similar results are obtained with the REM approach, which allows heterogeneity in the effect sizes between the different datasets.

In this manner, the microarray meta-analyses led to the identification of genes up- and down-regulated in each disease, and significant DEGs are selected as those displaying a FDR corrected p-value (q-value) < 0.05 (in both FEM and REM cases). Four other q-value cutoffs (0.005, 0.0005, 0.00005 and 0.000005) are also selected to validate our results on more stringent DEGs sets (see Appendix A). We only present here the cutoffs corresponding to FEM case, for simplify.

2.4.4 Comparisons of DEGs between the different diseases (Step 4)

Once the meta-analysis is done, for each cancer and CNS disorder sets of differentially expressed genes up- and down-regulated are obtained. Accordingly, each CNS disorder's DEGs are compared to each cancer type's DEGs. This is, the number of DEGs common between each CNS disorder and each cancer type are calculated (in two cases: up-regulated and down-regulated). The same strategy is done with each CNS disorder and cancer, with each of the control diseases.

Then the significances of the overlaps between the DEGs are assessed by a one-tailed Fisher's exact test. The Fisher's exact test assesses the difference between the data observed and the data expected, considering the given marginal and the assumptions of the model of independence. In this case, it calculates whether the genes in common between genes up-regulated in a certain kind of cancer and genes down-regulated in a CNS disorder are significantly different. The output of the Fisher's exact test is corrected for multiple testing by the Bonferroni approach. If multiple tests

of the same hypothesis are done, the chance of finding one or more of the tests to be positive increases. Thus, correction by false discovery rate (FDR) is fundamental and is used in multiple hypothesis testing to correct for multiple comparisons. The same procedure is applied for cancers, CNS disorders, and asthma, HIV, malaria, dystrophy and sarcoidosis (Figure 2.10). The background number of genes necessary for the Fisher's test is set to 14,538.

2.4.5 GSEA analyses (Step 5)

GSEA (Gene Set Enrichment Analysis) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes). For each CNS disorder and cancer type independently, a gene set enrichment analysis is undertaken using GSEA (Subramanian et al., 2005) on the output of the meta-analyses, and focusing on KEGG (Ramanan et al., 2012), Biocarta (Bio) and Reactome (Matthews et al., 2009) pathway databases.

Significant pathways are selected as those with q-value (FDR) < 0.05. Significant pathways in each disease are then compared to each others, and a network of pathways is built (Figure 2.13, Figures A.5 and A.6). For the KEGG pathways, further classification of the pathways in Metabolism, Genetic Information Processing, Cellular Processes, Environmental Processes and Organismal Processes, as provided by KEGG (keg) is done (Figure 2.14). Pathways corresponding to Human Diseases have been discarded.

2.5 Results

Following the scheme proposed (Figure 2.2), we present below the performance of the data selection described in Step 1 (Section 2.4.1). The outcome of the principal hypothesis regarding the *inverse comorbidity* between cancer and CNS disorders once data is preprocessed, and meta-analysis with FEM and REM strategies is also presented. We also show the result when the DEGs in cancer and pseudo-control disorders, and CNS and pseudo-control disorders are compared and analyzed respectively. The biological pathways enriched in the DEGs analyzed are also presented. At last, several potential candidates are proposed that could shed light on further research studies.

2.5.1 Microarray data selection through MetaQC

For each disease, the marginal impact of a meta-analysis on DEGs has been computed while studies are sequentially included into the meta-analysis. Figure 2.6 shows the number of DEGs for each disease, corrected by FDR = 0.05, when studies are added sequentially in the meta-analyses in the order of standardized mean rank (SMR) score. For each study, in x-axis the accumulation of the series of meta-analysis is represented. The order of addition follows the SMR score computed by MetaQC (Kang et al., 2012). Y-axis represents the number of DEGs detected. In all of them, the number of DEGs, under FDR=0.05, increases as more studies are added. Thus, including these additional studies to the meta-analysis provides an increased statistical power to detect more DEGs. This score is internally computed as $SMR = \frac{mean\ rank\ of\ all\ QC\ measures}{number\ of\ studies}$. In this way, this means that the selected data is adequate and suitable for a meta-analysis.

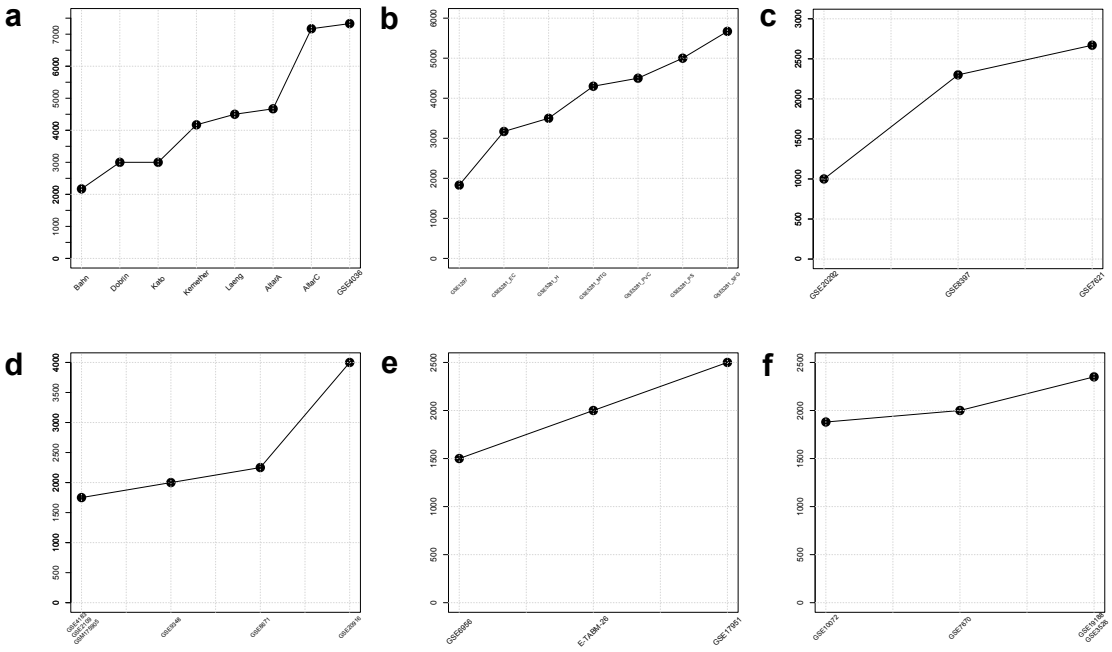


Figure 2.6: Marginal impacts on meta-analysis for DEGs detection in a) SCZ b) AD c) PD d) CRC e) PC and f) LC.

2.5.2 Significant *inverse comorbidity* between cancer and CNS disorders

As an example, we present here a gene expression profiling as a measurement of the activity (i.e., the expression) of thousands of genes at once, to create a global picture of how the gene expression is in each disease. Figure 2.7 shows the pattern of gene expression values when an Alzheimer's disease and a prostate cancer data sets are analyzed. Green cells indicate reduced expression, red cells are actively expressed. It can be observed that there are many parts within the plot having an inverse behavior between the gene expression in prostate data set (characterized by a red rectangle in the figure), and the gene expression in Alzheimer's disease (characterized by a blue rectangle in the figure).

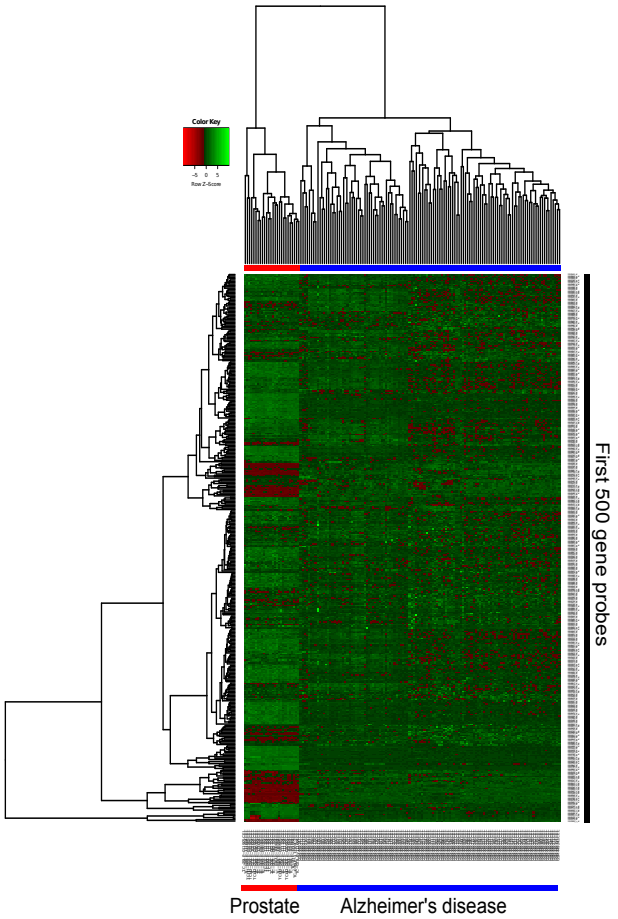


Figure 2.7: Heat map of the gene expression values in an Alzheimer's disease (in blue) and a prostate cancer (in red) data set.

For each CNS disorder and cancer type independently, we have undertaken meta-analyses from a large collection of microarray gene expression datasets to identify the genes that are significantly up- and down-regulated in disease when compared with their corresponding healthy control samples. Then, the DEGs of the CNS disorders and cancer types have been compared to each others.

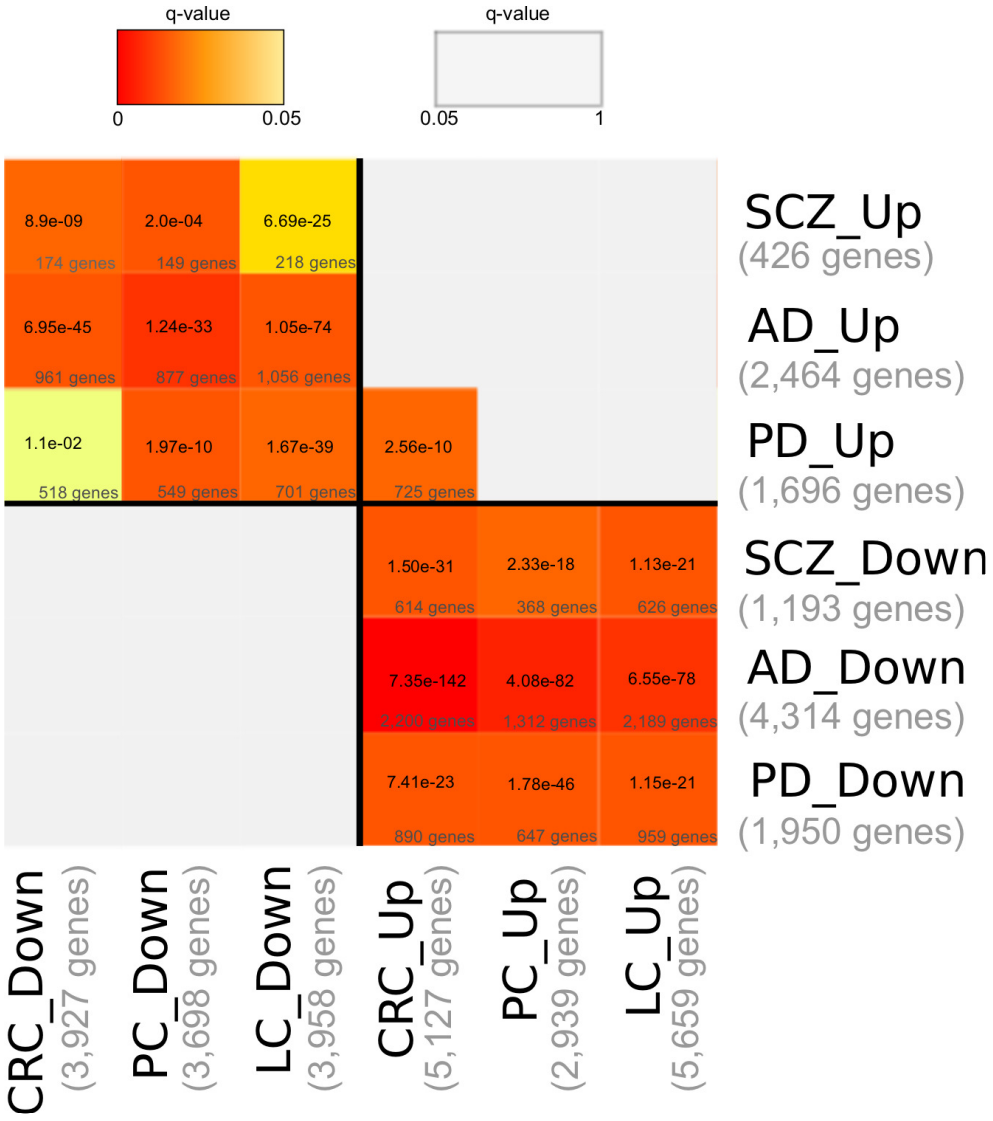


Figure 2.8: Comparisons of DEGs associated with CNS disorders and cancers with FEM.

Figures 2.8 and 2.9 contain the comparisons of DEGs associated with each pair CNS disorder–cancer, after having applied the data with FEM and REM respectively. Each cell includes the p-value of the significance of the DEGs overlap between the pair

of diseases. Cells are colored according to the significance of the overlaps (Fisher's exact test, Bonferroni correction for multiple testing, see Section 2.4.4). Grey cells correspond to non-significant overlaps (q-value > 0.05). The number of genes in common (overlapping genes) between the pair cancer–CNS disorder is included in each cell. Below the description of each disorder, the numbers of the total genes differentially expressed are also represented. There are significant overlaps between the DEGs up-regulated in CNS disorders and those down-regulated in cancers. Similarly, DEGs down-regulated in CNS disorders overlap significantly with DEGs up-regulated in cancers (Figures 2.8 and 2.9). It is also observed similar outcomes with the two different strategies, FEM and REM, even though with REM the results have lower significance values, but still very significant.

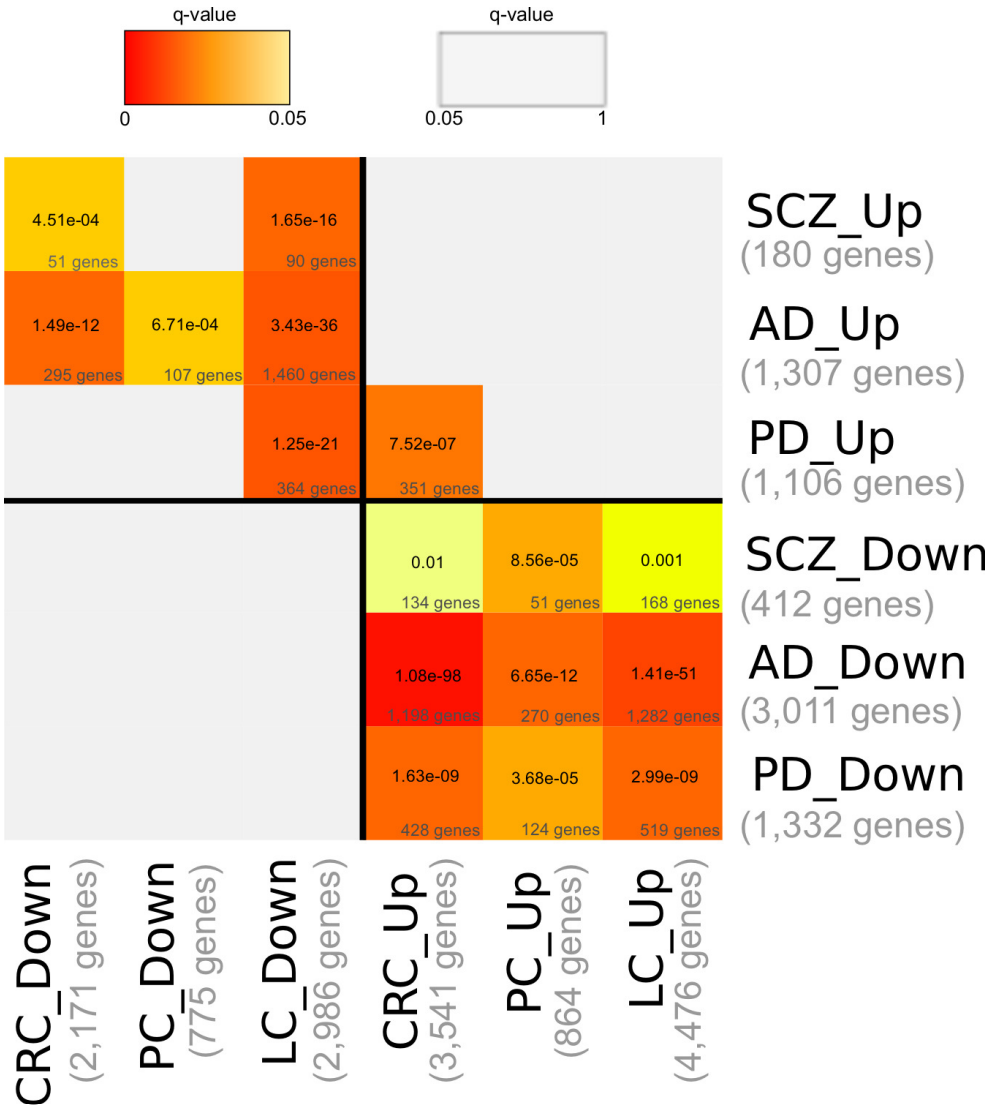


Figure 2.9: Comparisons of DEGs associated with CNS disorders and cancers with REM.

Significant overlaps between DEGs deregulated in opposite directions in CNS disorders and cancers are still observed while setting more stringent cutoffs for the detection of DEGs (q-values lower than 0.005, 0.0005, 0.00005 and 0.000005 by FEM, see Appendix A). A significant overlap between DEGs deregulated in the same direction is only identified in the case of CRC and PD upregulated genes with both methodologies (Figure 2.8).

2.5.3 Non-significant comorbidity between other diseases

Then, we have compared the CNS disorder and cancer DEGs with DEGs of a number of diseases for which, to our knowledge, *inverse comorbidities* have not been reported in the literature. These diseases, for which large enough expression datasets are available, included asthma, HIV, malaria, dystrophy and sarcoidosis.

Significant overlaps are observed between DEGs of all these diseases and DEGs of CNS disorders or cancers (Figures 2.10 and 2.11). However, patterns of expression deregulation in opposite directions, which are found to be characteristic of the relation between CNS disorders and cancers, are in most cases not observed with these other genetic or infectious diseases (Figures 2.10 and 2.11).

Figures 2.10a and 2.11a contain the comparison of DEGs associated with each pair cancer–control diseases with FEM and REM approaches respectively. Each cell includes the p-value of the significance of the DEGs overlap between the pair of diseases. Cells are colored according to the significance of the overlaps (Fisher's exact test, Bonferroni correction for multiple testing, see Section 2.4.4). Grey cells correspond to non-significant overlaps (q-value > 0.05). The number of genes in common overlapping between the pair cancer–other diseases, is included in each cell. Below the description of each disorder, the numbers of the total genes differentially expressed are also represented. Figures 2.10b and 2.11b contain in the same way, the comparisons of DEGs associated with each pair CNS disorder–control diseases.

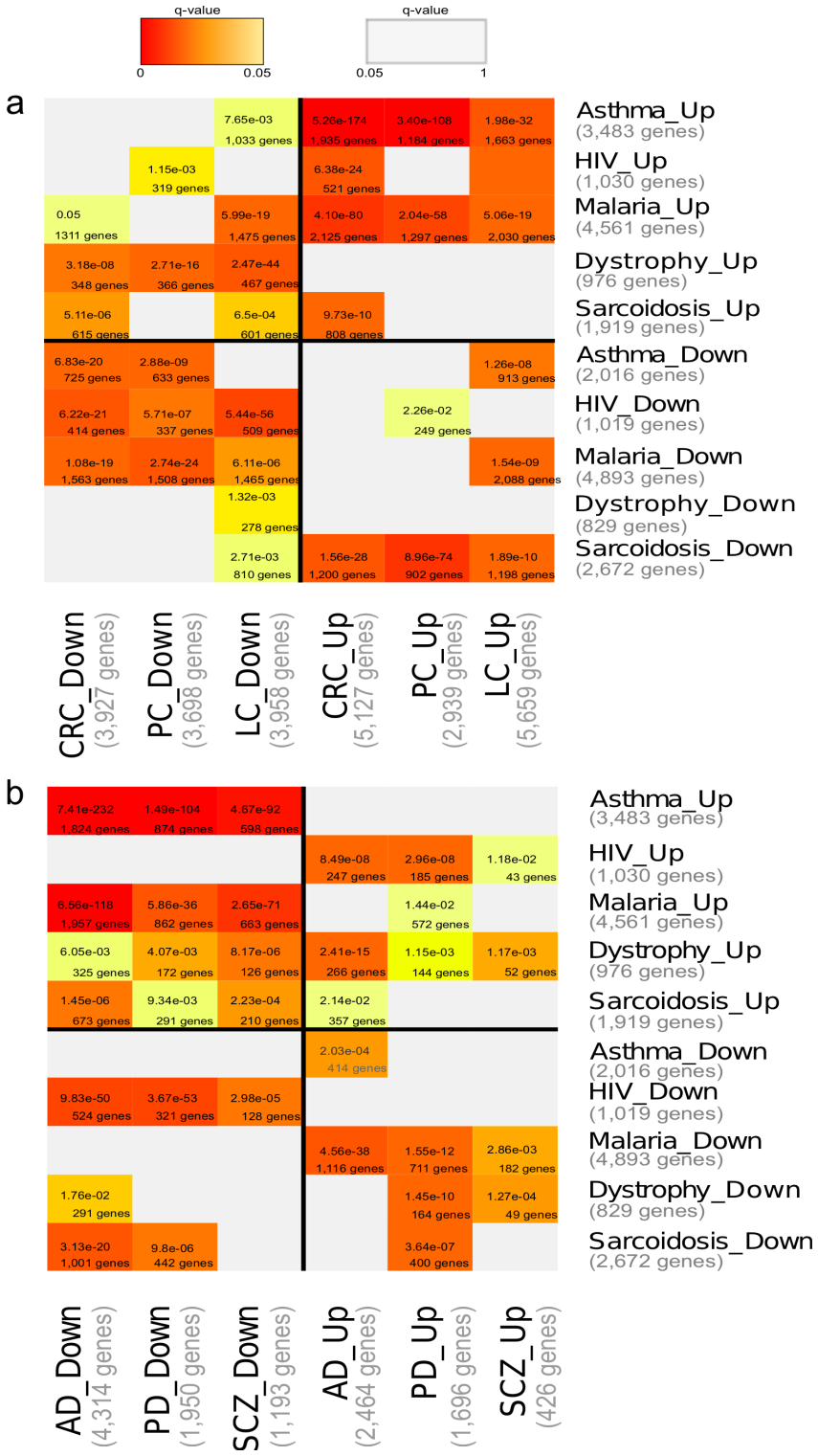


Figure 2.10: Comparisons of DEGs by FEM between a) cancer and asthma, HIV, malaria, dystrophy and sarcoidosis. And b) CNS disorders and asthma, HIV, malaria, dystrophy, and sarcoidosis.

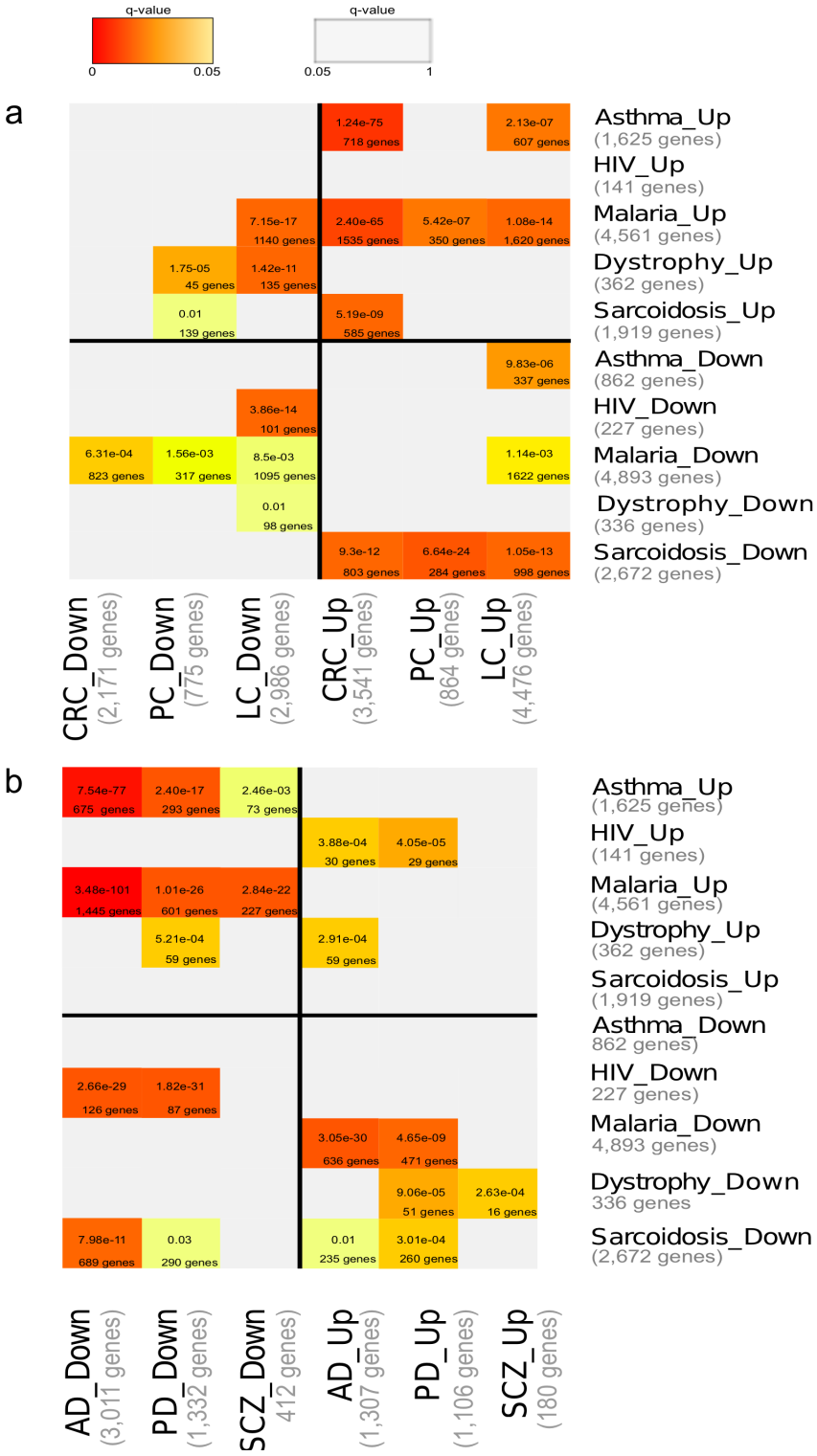


Figure 2.11: Comparisons of DEGs by REM between a) cancer and asthma, HIV, malaria, dystrophy and sarcoidosis. And b) CNS disorders and asthma, HIV, malaria, dystrophy and sarcoidosis.

An interesting observation is that the overlaps are predominantly significant between DEGs deregulated in the same directions, i.e., between up-regulated genes of the different diseases (or conversely between down-regulated genes), and could be a signature of putative positive comorbidities. It is to note that malaria and CNS disorders DEGs present overlaps between DEGs deregulated in opposite directions, contrarily to what is detected for other diseases. This observation will require additional research. Overall, these observations support the indication of a signature for *inverse comorbidity* in gene expression deregulations in opposite directions.

2.5.4 *PIN1* as putative candidate

The *PIN1* gene has been proposed previously as a putative link between the pathogeneses of AD and cancer (Behrens et al., 2009). Through the isomerization of a proline preceded by phosphorylated Ser/Thr residues, the PIN1 protein is known to be a key regulator of cell division (Lu, 2004). *PIN1* gene is typically over-expressed in human cancers and as such, it has been assessed as a potential target for anticancer drugs (Behrens et al., 2009). In addition, PIN1 is depleted in AD, it has been shown to restore the function of the phosphorylated tau protein, and mouse models in which this protein is knocked-down present neurodegenerative phenotypes (Lu, 2004; Liou et al., 2011).

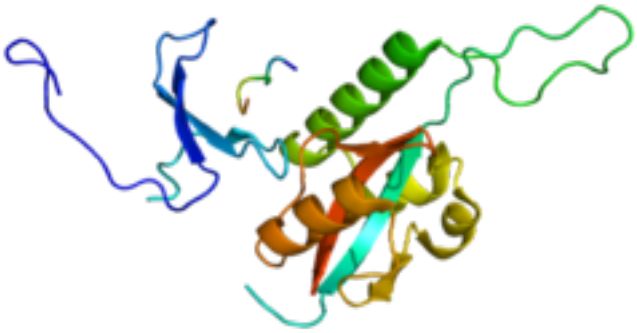


Figure 2.12: The PIN1 protein structure.

Our transcriptomic meta-analyses confirm and extend these observations as the expression of *PIN1* is down-regulated in AD and PD, and up-regulated in CRC. Another interesting case is the *ATP13A2* gene, involved in the intracellular cation homeostasis. *ATP13A2* is part of a list established by Devine et al. of familial PD genes frequently mutated in cancers (Devine et al., 2011). Indeed, loss-of-function mutations of *ATP13A2* have been associated with early-onset Parkinsonism, and somatic mutations have been independently observed in cancer (Devine et al., 2011). We identified *ATP13A2* as down-regulated in AD and PD, and up-regulated in the three cancer types considered.

2.5.5 Towards potential candidate links

In the light of these findings, our computational approach appears to be capable of identifying candidate genes potentially associated with *inverse comorbidity*. In particular, 74 genes may be of interest since they are simultaneously down-regulated in the three CNS disorders and up-regulated in the three cancer types examined (included in Table 2.3). RNA splicing (four genes: *PPIH, LSM4, NUDT21, SRSF2*) and aminoacyl t-RNA ligases (three genes: *FARSA, IARS, IARS2*) represent particularly interesting functions.

We also pinpoint two genes involved in lipid biogenesis (*ACLY* and *MECR*), and other two are transcription factors: *NME2* and *TFCP2*, for which a genetic association with AD is debated (Cousin et al., 2011). Finally, two other genes, *OAZ2* and the spermine synthase *SMS*, are dedicated to polyamine metabolic processes. Interestingly, defects in the spermine synthase gene are associated with the X-linked mental retardation Snyder-Robison syndrom (Cason et al., 2003), and spermine is often the most abundant polyamine in cancers (Huber and Poulin, 1995). The polyamine metabolic process hence may play a role in the pathological mechanisms of both CNS disorders and cancers.

Conversely, 19 genes are simultaneously upregulated in the three CNS disorders and downregulated in the three cancer types examined (included in Table 2.4), including for instance six genes involved in signal transduction (*TNFRSF1A, CDKN1A, NFKBIA, PTH1R, IL4R, MID1*). Particularly, *NFKBIA* is an interesting candidate because this gene is often deleted in glioblastoma (Bredel et al., 2011), although to our knowledge no mutations or polymorphisms have been described in CNS disorders.

PPIAP11, IARS, GGCT, NME2, GAPDHP1, CDC123, PSMD8, MRPS33, FIBP, OAZ2, IARS2, SLC35B1, APOO, TMEM189-UBE2V1, VDAC1, TMED3, SMS, , DNM1L, PRPS1, SRSF2, TMEM14D, TOMM70A, ATP6V1C1, NUP93, MRPL15, UBA5, PPIH, SMYD3, NIT2, SRD5A1, NUDT21, MRPL12, EEF1E1, MRPS7, TTPAL, BZW1P2, RP11-552M11.4, TSN, MECR, ZWINT, RPRD1A, UCHL5, NHP2P2, TFB2M, FEN1, CGREF1, IMPAD1, ARL1, , ACLY, MRPL42, LSM4, KPNA1, TIMM23B, RP11-164O23.5, RP11-762H8.2, FARSA, MRPL4, API5, RP3-425P12.4, RFC3, RANBP9, TFCP2, GMDS, CCNB1, TMEM177, GUF1, HSPA13, NMD3, GCFC2, TUBGCP5, TBCE, YKT6, PHF14, BRCC3

Table 2.3: DEGs significantly down-regulated in the three CNS disorders and up-regulated in the three cancer types (q-value < 0.05) (FEM approach).

MT2A, MT1X, NFKBIA, AC009469.1, DHRS3, CDKN1A, TNFRSF1A, CRYBG3, IL4R, MT1M, FAM107A, ITPKC, MID1, IL11RA, AHNAK, KAT2B, BCL2, PTH1R, NFASC

Table 2.4: DEGs significantly up-regulated in the three CNS disorders and down-regulated in the three cancer types (q-value < 0.05) (FEM approach).

2.5.6 Biological pathways in cancer and CNS disorders

In order to enhance the functional interpretation of the molecular bases of *inverse comorbidity*, we have broadened the comparisons of expression deregulations by considering pathways instead of individual genes (Ramanan et al., 2012).

We have identified the pathways that were significantly up- and down-regulated (see Section 2.4.5) in each of the six diseases independently. Among all the KEGG pathways (Kanehisa et al., 2008) significantly up- and down-regulated in the 6 diseases (CRC, PC, LC, SCZ, AD, and PD), 30 are shared by CNS disorders and cancers (i.e., significantly deregulated in at least 1 CNS disorder and 1 cancer type). Strikingly, of these 30 shared pathways, 24 (80%) are deregulated in opposite directions in CNS disorders and cancers (Figure 2.13, 63% and 86% for the Biocarta (Bio) and Reactome (Matthews et al., 2009) databases, respectively, Figures A.5 and A.6).

In Figure 2.13 KEGG pathways identified by GSEA which are significantly up- and down-regulated in each disease (cancer and CNS disorders) are represented in a network. The significant pathways have been compared between the 6 diseases and combined in a network representation. Node pie charts are colored according to the pathway status as cancer up-regulated (yellow), cancer down-regulated (blue), CNS disorder up-regulated (green), and CNS disorder down-regulated (red). The green–blue and yellow–red associations thus correspond to pathways deregulated in opposite directions in CNS disorders and cancers. Pathway labels are colored according to their classifications provided by KEGG as: Metabolism (green), Genetic Information Processing (yellow), Cellular Process (pink), Environmental Information Processing (red), and Organismal Systems (dark red).

The p53 signalling pathway is an anticipated candidate for deregulations in these diseases and for a role in *inverse comorbidity* (Behrens et al., 2009). Indeed, deregulations of the p53 signalling pathway are associated with the initiation and progression of cancers, while recent studies also point to a role for this pathway in CNS disorders (Tabarés-seisdedos and Rubenstein, 2013). As such, specific polymorphisms in the *TP53* gene are found in SCZ patients (Tabarés-seisdedos and Rubenstein, 2013). Although the *TP53* gene itself does not appear to be differentially regulated in our analysis, the p53 pathway is upregulated in CRC and LC, while it is downregulated in PD, AD and SCZ (Reactome database; Figure A.6).

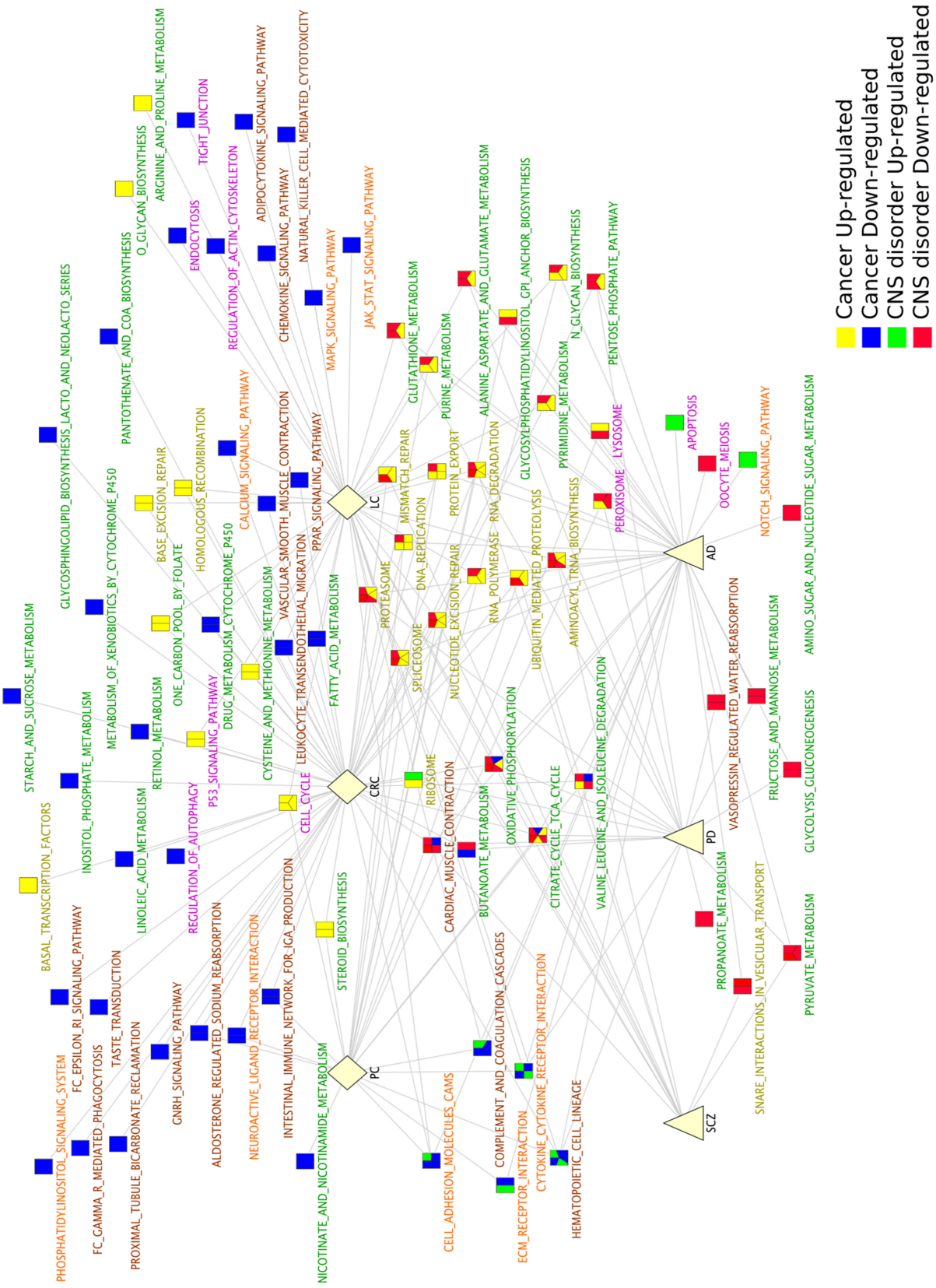


Figure 2.13: KEGG pathways significantly up- and down-regulated in each disease have been identified using the GSEA method (q-value<0.05).

Similarly, the Wnt pathway may be particularly relevant as mutations in the genes encoding APC and Beta-catenin, elements of the Wnt pathway, have been described in CRC, while beta-amyloid induced neurotoxicity in AD has been associated with impaired Wnt signalling (Lu, 2004; Behrens et al., 2009). Furthermore, alterations in the Wnt signalling pathway are known to be involved in SCZ (Okerlund and Cheyette, 2011). In our meta-analyses, we have found the Wnt pathway to be downregulated in AD and PD, and upregulated in CRC (Reactome database; Figure A.6).

Aside the Wnt and p53 pathways, our analysis reveals other pathways related to protein folding and protein degradation displaying patterns of down-regulation in CNS disorders and up-regulation in cancers, and that may be relevant for *inverse comorbidity*. For instance, the Ubiquitin/Proteasome system is consistently downregulated in CNS disorders and upregulated in cancers according to the three pathway databases analyzed (Figure 2.13, Figures A.5 and A.6). The inverse relationship between the levels of expression deregulations of these pathways possibly suggests opposite roles in CNS disorders and cancers.

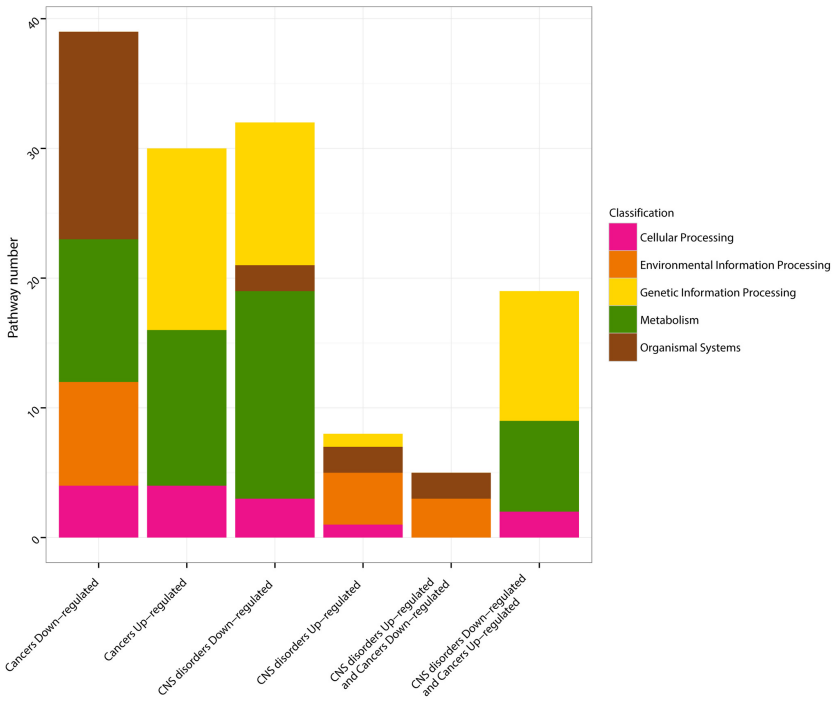


Figure 2.14: The KEGG pathways identified by the GSEA analysis (FEM).

Figure 2.14 shows the descriptions of the KEGG pathways identified by GSEA analysis as significantly up- and down-regulated in CNS disorders, in cancers, and simultaneously up-regulated in CNS disorders/down-regulated in cancers, and down-regulated in CNS disorders/up-regulated in cancers (q-values < 0.05, Figure 2.13). They are classified as Metabolism (in green), Genetic Information Processing (in yellow), Cellular Process (in pink), Environmental Information Processing (in orange) and Organismal Systems (in brown), according to the classification provided by KEGG

This detailed examination of the KEGG pathways deregulated in opposite directions in CNS disorders and cancers finally reveale that 89% of the KEGG pathways that are up-regulated in cancers and down-regulated in CNS disorders are related to Metabolism and Genetic Information Processing (Figure 2.13, Figure 2.14). By contrast, the pathways down-regulated in cancers and up-regulated in CNS disorders are related to the cell's communication with its environment (Environmental Information Processing and Organismal System; Figure 2.13, Figure 2.14). Hence, global regulations of cellular activity may account for a protective effect between inversely comorbid diseases.

2.6 Conclusions

We have developed a computational approach in order to compute transcriptomic meta-analyses in complex disorders. Particularly, between some types of cancer and CNS disorders in the context of the *inverse comorbidity*, since a genetic base exists. This complex biomedical problem has been studied at the medical level, based on solid population and epidemiological studies, but there is little information at the molecular level. We here advance for the first time evidences of the genetic base associated with *inverse comorbidity* at the molecular level, analyzing a relevant aspect within Genetics, i.e. gene expression. A novel, remarkable, and significant overlap is observed between the genes up-regulated in cancers and down-regulated in CNS disorders, as well as between the genes down-regulated in cancers and up-regulated in CNS disorders. These gene expression deregulations in opposite directions are also observed at the level of pathways, and point to specific genes and functions the deregulation of which could promote CNS disorders and simultaneously lowers the initiation or progression of cancer.

A molecular interpretation of the *inverse comorbidity* between CNS disorders and cancers could be that the down-regulation of certain genes would at the same time increase the risk of developing CNS disorders, while reducing the risk of developing cancer. The up-regulation of other genes would reduce the risk of developing CNS disorders and increase the risk of developing cancer. But we might be cautious, because there are several real limitations: the gene expression data analyzed is derived from different tissues, and from different patients. For instance, we do not know whether a gene up-regulated in Alzheimer's disease is down-regulated in the corresponding lung tissue in the same individual, and this is the reason for not developing cancer (and the other way around). The cases and controls data from the neurological disorders correspond to different individuals, since it is not possible to take a normal brain tissue from the patient under study. Indeed, the gene expression may vary across the different parts of the brain. Moreover, in cancer data, normal and tumoral cells are mixed and even though control tissue data are used, part of the information might be lost. There is a large variability within the individuals and the diseases under study, they are not homogeneous since there are many cancer subtypes and in such a manner, many differential behavior subtypes. Thus, not all the patients suffering a disease will have the same grade of protection.

New data and further analyses will be necessary to conclude to a direct protective effect of gene expression deregulations in cancer-prone tissues of patients suffering from CNS disorders. Indeed, the DEGs analyzed in this study are computed for each disease in the corresponding affected tissues, and cannot be extrapolated to gene expression deregulations in other tissues of the same patients. However, despite these limitations, the identification of antagonistically deregulated genes and pathways in complex diseases that have been previously described as inversely comorbid provides, to our knowledge, the first systematic insights into the possible molecular basis of these associations. The principal result we here present tries to open a new perspective.

It suggests that the up-regulation of a set of genes or processes could increase the incidence of CNS disorders and simultaneously lower the chances of developing cancers, while the down-regulation of another set of genes or processes could contribute to a decrease in the incidence of CNS disorders while increasing the cancer risks. The individuals delivering post-mortem brain samples in CNS disorders, or tumor tissues in the case of cancers, are likely to have received drug treatments. Hence, the observed

expression deregulations could be the consequence of the drugs administered to the patients. If this is the case, it can be hypothesized that some of the drugs used to treat CNS disorders might be able to revert the expression of a number of cancer genes. In this context, the repurposing of drugs from the CNS disorders to the cancer field could open new therapeutic avenues. Indeed some punctual observations have been made. For example, the thioridazine, an anti-psychotic drug antagonizing the dopamine receptor and potentially able to alter physiological states and expression patterns, have been reported to target cancer stem cells selectively (Sachlos et al., 2012). Another example is a recent work published by Yale School of Medicine researchers reported in the journal Annals of Neurology. They have proved that a failed drug on treating solid tumor appears to restore synaptic connections and reduced inflammation, and the animal's memory, a hallmark of Alzheimer's disease. In this manner, memory and the connections between brain cells are restored in mice with a model of Alzheimer's given this experimental cancer drug (Kaufman et al., 2015). Despite these two last observations, the effect of the drugs cannot explain by themselves the observed *inverse comorbidity*. For instance, several works have noted that the relatives of patients suffering schizophrenia have less probability of developing any cancer (Gal et al., 2012; Ji et al., 2013), suggesting that genes associated with schizophrenia might confer reduced cancer susceptibility. There is a genomic and molecular base that determines the behavior of the individual in general, and the effect of the drug in each one, in particular. More data is necessary to be able to establish a relationship between some drugs and the genetics behind the *inverse comorbidity*, and the observed gene expression patterns.

From the computational and statistical point of view, similar results have been obtained with both FEM and REM approaches. They do not only use the effect size, the variance is essential when studying the variability within the study and across the datasets. However, meta-analytic strategies such as Fisher, use only p-values and they do not take into account the variance for instance. FEM and REM are very referenced as the most appropriate to do a microarray meta-analysis. The use of combination of p-values (Fisher) would be suitable when gene expression raw data was available, but a list of genes with the associated p-values (usual in the published works).

Finally, the analyses of inverse expression deregulations could serve as a new approach to investigate relations between complex diseases, of which the ones reported here between CNS disorders and cancers can be considered as an initial example.

3 Study of the stability of protein interaction networks in cancer and CNS disorders

3.1 Summary

Molecular networks provide a powerful tool for the study of biomedical systems, in particular several studies have detected alterations of the network structure associated to disease states. Here we propose that diseases cannot only alter the structure of the network but also its stability. To evaluate network stability we have developed a new methodological framework. Our approach takes advantage from the classical deterministic Simulated Annealing algorithm to work with discrete states. Adjusted energy values are used to compare the network stability in disease and control states. The results show that cancer networks are less stable than the Alzheimer's disease ones. These results can be interpreted in terms of our previous observations on cancer and Alzheimer's disease *inverse comorbidity*, i.e., Alzheimer's disease patients have lower than expected risk to suffer cancer.

3.2 Introduction

CNS or neurological disorders and cancer are two current global health priorities. Interestingly, epidemiological evidence is mounting that patients with certain neurological disorders, including those suffering from schizophrenia (SCZ) and Alzheimer's disease (AD), have a lower than expected tendency to develop some forms of cancer (Behrens et al., 2009; Tabarés-Seisdedos et al., 2011; Behrens et al., 2012; Tabarés-seisdedos and Rubenstein, 2013). Hence, we performed a systematic meta-analysis of

gene expression in order to investigate the molecular mechanisms that might underlie such *inverse comorbidity*, identifying genes and pathways differentially expressed in neurological disorders and some types of cancer (Ibáñez et al., 2014). Interestingly, we found a common set of genes and biological processes that were apparently deregulated in opposing directions in cancers and neurological disorders.

Here, we set out to broaden our understanding of the molecular basis underlying the differences between cancers and neurological conditions. As such, and given that the central dogma of molecular biology dictates that information flows from genes to proteins via RNA (Crick, 1970) (see Figure 1.3), we have integrated gene expression data with Protein-Protein Interaction Networks (PPINs) in order to study these differences in terms of network organization rather than at the level of individual genes. Gene expression data informs whether a gene that encodes a given protein is active or not. Yet proteins function in the context of their interactions with other proteins, interactions that are described in PPINs in which each protein represents a node in the network. The strategy of combining PPINs with gene expression is artificial, since the networks that are available correspond to a particular state rather than diseases, individuals or conditions.

In PPINs, it is assumed that proteins corresponding to genes that are not active (i.e. unexpressed) will not interact with their potential partners. Therefore, the production of RNA by genes is commonly used as a proxy of the activity of the gene, and this is correlated with the activation of molecular systems within PPINs that underlie physiological and developmental processes. Indeed, in many cases, deregulation of gene expression provokes dramatic phenotypic changes, as occurs in several diseases (Kaern et al., 2005).

Protein interaction maps have been used to study the molecular organization of cellular systems and the perturbations in them created by disease. PPINs reflect the functionality of interacting proteins and for example, the consequence of a single gene deletion in the yeast *Saccharomyces cerevisiae* would appear to depend on the position of the gene product within the PPIN (Jeong et al., 2001). Thus, the most important proteins for a cell's survival are highly connected (Jeong et al., 2001; Wuchty and Almaas, 2005), and altering them has profound effects on the PPIN. In terms of cancer, it is thought that cancer related proteins correspond to central hubs and that they are highly connected within networks (Jonsson and Bates, 2006). Indeed, the genomic

and network characteristics of genes mutated in cancer seem to confirm that these genes tend to encode central hubs within PPINs (Rambaldi et al., 2008). In addition, PPINs have been used as background layers when mapping gene expression data in order to gain information about the state of the nodes and their possible dynamics (Chuang et al., 2007; Pujana et al., 2007; Hudson et al., 2009; Milanesi et al., 2009; Komurov and Ram, 2010; Schramm et al., 2010; Teschendorff and Severini, 2010; West et al., 2012; Börnigen et al., 2013; Liu et al., 2013b; van Pel et al., 2013). For example, genes that are over expressed in lung cancer are more strongly connected than those that are suppressed or selected at random (Wachi et al., 2005).

We here hypothesize that PPINs related to cancer are more unstable than those based on neurological data. This may be because there are more active interactions between cancer related proteins and thus, a mutation or change in any of these would cause an important destabilization of the network. By contrast, proteins corresponding to genes affected in neurological disorders have less active connections and consequently, they are less susceptible to destabilization. In this context, we present an approach based on the combination of gene expression data and PPINs to study the relationship between cancers and neurological disorders. To achieve this, we associate each protein (or node) in the network with a state that is directly related to the level of expression of the corresponding gene. The expression data used is derived from a large series of experiments carried out on cancer and neurological disorders in humans, information that makes the PPINs disease specific, and that allows comparative studies to be performed.

In terms of the computational methodology to study the differences between disease specific networks, we have found an appropriate equivalence in the deterministic Simulated Annealing (DSA) algorithm proposed previously (Duda et al., 2001). The DSA algorithm was designed to find the optimal solution inspired by different biological or physical phenomena. The DSA is based on the shifting of metals from an unstable state as a liquid to a stable solid state, a process mediated by a decrease in the temperature of the material. These transformations can be simulated by the evolution of the states of interconnected network nodes that evolve until an optimal solution with minimal energy is reached. This evolution is controlled through an energy minimization process that determines the network's stability as the energy decreases. Therefore, lower energies correspond to greater stability.

Inspired by the DSA algorithm, we have designed and implemented a new method to measure the stability of PPINs based on an energy function. In this approach the concept of stability differs from that in the original DSA, in which the network evolves towards states with different stabilities via temporal transitions or another equivalent value (Pajares and de la Cruz, 2004; Sánchez-Lladó et al., 2011). The proposed approach used in this study computes the energy based on existing interactions and it computes the energy difference between two states, such as disease and control samples. In this manner, the temporal aspect of the original DSA is reduced to the comparison between a reference and a new model. The reference state can be considered to be equivalent to the initial state and the new model as a single progressive step. Furthermore, any simulated annealing process (DSA or probabilistic) is driven by an optimization process in order to achieve stable states (minimum energy values). By contrast, since only one transition is considered in our approach, there is no optimization process involved and local minima energy are avoided. These substantial differences from the original DSA are introduced to make it possible to perform a large scale systematic comparison of networks associated to cancer, neurological disorders and normal controls for which the information available comes from experiments carried out at only one time point, representing a single state of these conditions.

3.3 Materials

The protein interaction and gene expression data used in this study are obtained from PPIN and gene expression data sets.

3.3.1 The protein–protein interaction network

We have used the human PPIN from the Protein Interaction Network Analysis database (pin) version October 2011 (Wu et al., 2009). PINA is an integrated platform of PPIN data that has been extracted from six different public databases: IntAct, MINT, BioGRID, DIP, HPRD, and MIPS/MPact. It includes self-interactions, interactions predicted by computational methods, and interactions between human proteins and proteins from other species. Moreover, it has recently been used in other similar studies (Xia et al., 2011), in which they have compared protein interactions network characteristics of cancer proteins to see whether cancer proteins interact strongly

within the human PPIN.

Besides the PINA network, we also have used two additional PPINs in order to guarantee that a similar outcome is obtained: The Human Protein Reference Database (HPRD), version April 2010 (hpr) that contains pairs of human protein interactions based on experimental evidence from the literature and that has been used in several studies (Teschendorff and Severini, 2010; West et al., 2012), and the Human Integrated Protein-Protein Interaction rEference (HIPPIE) version September 2014 (hip) that incorporates a human PPI dataset with a normalized scoring scheme, integrating data from HPRD, BioGRID, IntAct, MINT, Rual05, Lim06, Bell09, Stelzl05, DIP, BIND, Colland04, Lehner04, Albers05, MIPS, Venkatesan09, Kaltenbach07, and Nakayama02. We have selected the interactions from these PPINs with a curated score above 0.73 in order to be confident that the pairs of proteins interact (Schaefer et al., 2012).

3.3.2 Gene expression data sets

Measuring gene expression with microarrays is now a common molecular biology approach in biomedicine, making it possible to simultaneously measure the relative expression of thousands of genes under different experimental conditions (mit, 2002). Thousands of gene expression data sets are available in public databases, each containing a description of the corresponding biomedical origin of the sample, the analytic procedures followed, and the experimental results in terms of expression (i.e., the amount of RNA produced for each gene in the genome).

Raw experimental gene expression data (CEL files) for ovarian, colon, liver and kidney datasets have been downloaded from the Barcode human transcriptome repository (bar), and for the SCZ and AD datasets they have been downloaded from the NCBI GEO omnibus (geo) and Stanley Medical Research Institute Online Genomics Database (smr) (SMRI). Importantly, each dataset corresponds to a collection of disease and control samples. For the analysis we have filtered out the cases with too few disease/control cases (less than 9) and we only have used those produced in the same platform (Affymetrix array GeneChip Human Genome U133 Plus 2.0), rendering information on 23,945 human genes. This technical platform has been widely used, and using the same platform on all data sets facilitates comparative studies and ameliorates potential experimental errors. Table 3.1 summarizes the data included in this study. The platform field indicates the name of the microarray platform used in

the gene expression dataset, the sample size the total number of patients (cases and controls) included in each dataset, and the source the identifier of the experimental dataset in GEO and SMRI repositories.

Table 3.1: Gene expression datasets.

Tissue	Platform	Sample Size	Source
Alzheimer's disease			
Entorhinal Cortex	HG-U133Plus2	23	GSE5281
Hippocampus	HG-U133Plus2	23	GSE5281
Medial Temp. Gyrus	HG-U133Plus2	28	GSE5281
Posterior Singulate	HG-U133Plus2	22	GSE5281
Primary Visual Cortex	HG-U133Plus2	31	GSE5281
Superior Frontal Gyrus	HG-U133Plus2	34	GSE5281
Hippocampus	HG-U133Plus2	24	GSE1297
Schizophrenia			
Postmortem cerebellum	HG-U133Plus2	28	GSE4036
Postmortem frontalBA46 cortice	HG-U133Plus2	55	Dobrin
Postmortem hippocampus (CA1)	HG-U133Plus2	41	Laeng
Postmortem thalamus (MD)	HG-U133Plus2	26	Kemether
Colorectal cancer			
Colon tissue	HG-U133Plus2	33	GSE4183 GSE7307 (GSM175905)
Colon tissue - stages	HG-U133Plus2	43	GSE2109 GSE17537
Ovarian cancer			
Ovary tissue	HG-U133Plus2	278	GSE9891 GSE2109 GSE9890 GSE3526 GSE7307

Table 3.1 – Gene Expression datasets

Tissue	Platform	Sample Size	Source
Liver cancer			
Liver tissue	HG-U133Plus2	129	GSE6222 GSE9829 GSE9843 GSE6956 GSE11045 GSE7307
Liver tissue - stages	HG-U133Plus2	8	GSE6764
Kidney cancer			
Kidney tissue	HG-U133Plus2	119	GSE2109 GSE11151 GSE12090 GSE8050 GSE9489 GSE9493 GSE11045 GSE11151

For each disease, several datasets are included. For AD and schizophrenia we have included gene expression studies from different parts of the brain, because the expression of the genes could vary depending on the region of the brain. For colorectal and liver cancer, additional datasets have been included to study the stability of the PPINs during the cancer evolution (see Section 3.5.4).

3.4 Methods

In order to study the stability of the PPIN in cancer, neurological and normal samples, we have implemented an original method inspired by the well-known DSA approach that has been customized to study neighbor-energy (*nE*). In this case, stabil-

ity describes a network state that is not significantly altered, even when fundamental properties have changed or perturbations have been introduced. From the biological point of view, network instability could reflect a situation where mutations in a key protein involved in many interactions will alter several associated biological processes.

A filtered PPIN (Section 3.4.1), and preprocessed and normalized gene expression data (Section 3.4.3) for three different conditions (cancer, normal and neurological disorders), are the inputs for our approach (Section 3.4.4). A scheme of the workflow is presented in Figure 3.1, where preprocessing and filtering are clearly represented as two separate modules. The first four boxes represent the selection, inclusion and preparation of the data (gene expression and PPIN) that correspond to the input of our proposing approach.

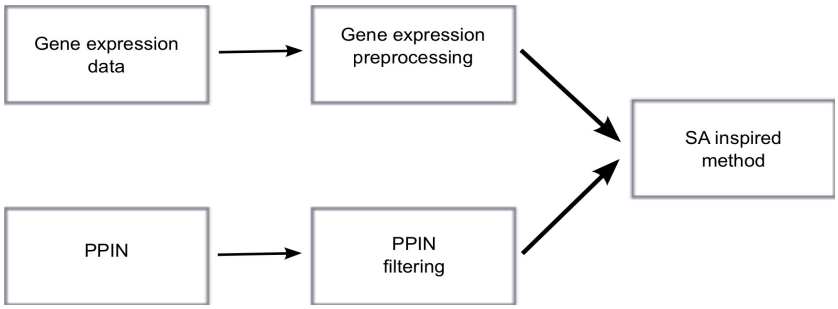


Figure 3.1: Proposed flow chart.

3.4.1 Protein–protein interaction network filtering

Data from the PINA network are filtered by requiring experimental evidence for PPIs, removing redundancy and self-interactions, as well as interactions involving proteins that are not from *Homo sapiens*. Thus, we only consider those interactions between proteins that are also detected in the Human Genome U133 Plus 2.0 microarray platform. The resulting filtered PINA network contains 10,650 proteins with 63,119 interactions. In Figure 3.2a a subnetwork of the filtered PPIN is shown.

3.4.2 Sub-network related to synaptic vesicle cycle

A sub-network of proteins encoded by genes related to the synaptic vesicle cycle is analyzed, retrieving proteins in the synaptic vesicle (SV) cycle from the KEGG pathway (keg) (hsa04721 pathway id, version September 2014). The number of genes involved in the SV cycle pathway are 63, and 50 out of 63 genes are detected in microarrays. The resulting sub-network contains 50 proteins and 3,815 interactions.

3.4.3 Microarray gene expression preprocessing

Handling microarrays requires the preprocessing of each individual microarray to estimate the expression of each gene in the array. Gene expression data from ovarian, colon, liver and kidney cancers, and from SCZ and AD samples, are normalized by frozen Robust Multiarray Analysis (fRMA) (McCall et al., 2012) from the R Affy package (Gautier et al., 2004). Background-corrected gene intensities are obtained by applying fRMA processes to each array individually, and accounting for probe variability, batch effects, probe effects, array-to-array variability, and background noise. The samples are then processed using Barcode (McCall et al., 2012) in order to convert gene intensities into estimates of gene expression (Z-score, Figure 3.2b). Additionally, gene intensities are mapped into a binary vector of "ones" and "zeros" that denote whether a gene is expressed (1, when the Z-score is higher than a threshold value: 4,98 by default) or not (0) in each sample (Figure 3.2b and Table 3.1: (McCall et al., 2011; Zilliox and Irizarry, 2007). These values are used in Equation 3.1, in which it is necessary to specify whether a gene is expressed or not.

To compare the Z-score between these diseases, we have normalized them using the $pnorm$ function of the R stats package to calculate the normal distribution function of each Z-score. This normalization step is commonly employed to avoid values in a given range dominating other values. High Z-scores indicate intense gene expression, while small Z-scores correspond to weak expression. For expressed genes, defining S as the normalized Z-score, $S = pnorm(Z-score)$, represents the probability of the gene being expressed. When the gene is not expressed, $S = 1 - pnorm(Z-score)$ indicates the probability of the gene not being expressed. These S values are used in Equation 3.2. Hence, each state in the system would represent the significance (S) of the expression of each gene (Figure 3.2b). In summary, for each disease we have

associated a binary value reflecting whether the gene is expressed or not (one or zero, respectively), attributing a value and a significance to the expression of each of the 10,650 genes in the network (Figure 3.2b).

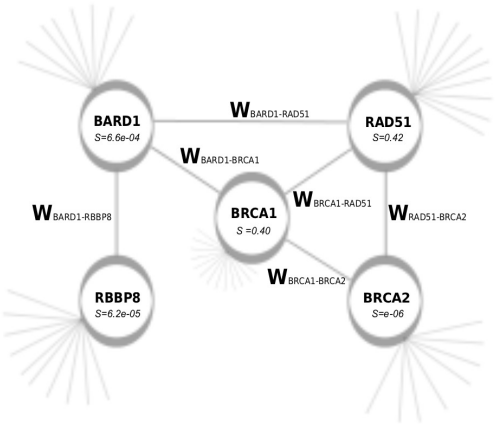
a Protein interactions

GeneA	GeneB
BRCA1	BRCA2
BRCA2	RAD51
RAD51	BRCA1
RAD51	BARD1
BARD1	BRCA1
RBBP8	BARD1
BRCA1	RBBP8
CHEK2	STRAP
GH1	BTRC
WNK1	MAP3K2
CLTC	STAMBPL1
MLLT10	DOT1L
ESR1	CUL4B
FLOT1	ARG1
NFKB2	DYNC1H1
.	.
.	.
.	.
FOS	RNF20
ERBB2	NRG1
MYC	EP400

b Gene expression

Gene	Z-score	Normalized Z-score (S)	Barcode
DDR1	8.24	1	1
RFC2	1.49	0.07	0
PAX8	6.79	0.97	1
GUCA1A	-0.66	0.6	0
UBA7	6.86	0.98	1
THRA	-0.01	0.5	0
PTPN21	0.43	0.33	0
CCL5	5.02	0.99	1
BRCA1	-0.23	0.40	0
BRCA2	4.56	2.5e-06	0
RAD51	0.19	0.42	0
BARD1	3.21	6.6e-04	0
RBBP8	3.84	6.2e-05	0
MAPK1	4.65	1.7e-06	0
VPS18	0.24	0.4	0
.	.	.	.
.	.	.	.
.	.	.	.
PCP2	-1.26	1	0
MTCP1	-0.06	0.62	0
NAF1	2.78	2.7e-03	0
ATP6V0C	12.10	1	1

c Subnetwork example



$$\textbf{partial nE} = -\sum W_{BARD1\text{-}RBBP8} * S_{(RBBP8)} * S_{(BARD1)}, W_{BARD1\text{-}BRCA1} * S_{(BARD1)} * S_{(BRCA1)}, W_{BARD1\text{-}RAD51} * S_{(BARD1)} * S_{(RAD51)}, W_{RAD51} * S_{(RAD51)} * S_{(BRCA1)}, W_{RAD51\text{-}BRCA2} * S_{(BRCA2)} * S_{(RAD51)}), (W_{BRCA1\text{-}BRCA2} * S_{(BRCA1)} * S_{(BRCA2)}$$

*partial nE = - (+1*6.2e-05*6.6e-04 +1*6.6e-04*0.40 + 1*6.6e-04*0.42 - 1*0.42*0.40 + 1*2.5e-06*0.42 + 1*0.40*2.5e-06*

Figure 3.2: (a) Filtered PPI subnetwork. (b) Preprocessed and normalized gene expression data. (c) Representation of the nodes in a PPIN sub-network and an application of the algorithm to a PPIN sub-network.

3.4.4 Approach inspired by simulated annealing algorithm

To study network stability we have adopted an approach based on the SA concept, a probabilistic method that allows the global minimum of a generic cost function to be found (S. Kirkpatrick et al., 1983; Cerny, 1985). This procedure reproduces the way the structure of a solid reaches its minimum energy configuration through cooling, becoming "frozen" at this minimum energy.

A full description of the DSA is included in Appendix B (Haykin, 1994; Duda et al., 2001), which also follows a physical analogy based on a set of interconnected nodes, each one with its associated state. During the cooling process forces between interconnected nodes act on the structure, which evolves until each node reaches a stable state. Thus, the nodes interacting with other nodes within the system influence one another with a defined weight.

Our algorithm is inspired on the definition of a *nE* function that measures the stability of the network, as well as on the general deterministic approach whereby a lower *nE* is related to greater stability. In our case, using a *nE* function that decreases in function of the interactions or over time does not make sense given the characteristics of the biological problem. Indeed, our approach does not evolve through iterations or time and thus, this part of the algorithm was not considered.

Our system is represented by a PPIN in which nodes represent proteins associated to the expression of the corresponding gene (S_i describes the significance (S) of a gene i being expressed or not). Our approach is applied to estimate the dynamic structures in the PPIN (Figure 3.2c), where S_i represents the state of the node in the original DSA approach, and the edges reflect the interactions existing between proteins. Each W_{ij} represents the weight required (Equation 3.1), where W_{ij} is inversely associated to the existence of the interaction between two proteins. If the two genes i and j are both expressed, then the two corresponding proteins can interact (W_{ij} value -1). The value of W_{ij} will be $+1$ if the interaction is not possible because one of the two genes

is not expressed.

$$W_{ij} = \begin{cases} -1 & \text{if i expressed, j expressed} \\ +1 & \text{if i or j not expressed} \\ +1 & \text{if i not expressed, j not expressed} \end{cases} \tag{3.1}$$

Consistent with the main idea of the algorithm, the *local_nE* is defined as the sum of the energy from all the nodes connected to a given node i. This influence is calculated by multiplying the expression of each gene (normalized value of expression, S) by the associated weights of the connected nodes (W_{ij}), as summarized in Equation 3.2.

$$local\_nE(i) = -\sum_j W_{ij} * S_i * S_j \tag{3.2}$$

According to the definition in Equation 3.2, the *local_nE* is maximal when $W_{ij} * S_i * S_j$ is at its minimum, representing active connections between nodes of expressed genes (Equation 3.1, case 1) and indicating that any alteration in this node will destabilize the network. The value of the *local_nE* decreases for those node connections that involve at least one gene that is not expressed in that condition, reflecting the fact that the interactions cannot take place (Equation 3.1, cases 2 and 3). In these situations, the *local_nE* achieves its minimum values indicating network stability.

The *local_nE* function measures the stability of a single protein or node i in function of its neighborhood, i.e., only with respect to the directly interacting partners, and not within the entire network. The global *nE* value (Equation 3.3), and therefore the stability of the entire network, will be a consequence of the equilibrium between interactions among active (corresponding to the expressed genes) and inactive nodes (corresponding to the non-expressed genes).

$$nE = \sum_i local\_nE(i) \tag{3.3}$$

3.4.5 Computation of network robustness

To assess the robustness of the system, we have analyzed how the network structure changes as nodes are removed in accordance with previously defined procedures (Iyer et al., 2013). Changes in the network structure are evaluated in terms of the size of the largest connected component of the network. Networks in which the largest component decreases faster than that of the original network are considered to be less robust to perturbations. Thus, nodes are removed in decreasing order of their *local_nE* scores (Equation 3.2), removing the proteins (or nodes) with higher *local_nE* values first (i.e., those with more active connections), and those with the lowest *local_nE* scores last (i.e., those less connected to their neighbors).

Network robustness is measured through the *R-index* in Equation 3.4, where α corresponds to the size of the largest connected component within the network after a node is removed.

$$R = \frac{1}{N}\sum_{i=1}^{N}\alpha(i/N) \tag{3.4}$$

We have computed the *R-index* for cancer and normal control samples at each step after the removal of nodes in function of the order of *local_nE* scores.

3.5 Results

Using this new approach, we have analyzed four gene expression datasets for cancer (ovarian, colon, liver and kidney), four data sets for SCZ, and five for AD (Table 3.1) each having sufficient disease and control samples, and fulfilling our quality control criteria (see Section 3.3.2). For each disease and data set, PPIN stability is assessed in both the disease and control samples. In other words, we have simulated a weighted interaction network for each sample, mapping *S* into the PPIN, directly applying the proposed algorithm and obtaining a *nE* value. The distribution of the *nE* values for the normal (N) and disease (C) conditions are then studied (Figure 3.3) and a global *nE* is obtained for each disease.

3.5.1 Increased neighbor-energy in cancer tissue

The cancer PPINs present characteristic instability, reflected by higher *nE* values than their normal control samples (Figure 3.3a). A Mann-Whitney (Wilcoxon-rank) test is used to evaluate whether the medians of a test variable differed significantly between the normal and cancer samples, which proves to be the case for each tissue (represented below the x-axis). The Wilcoxon Mann-Whitney test is a non-parametric statistical test that computes the difference between the distributions of data collected in two experimental conditions (control and cases in our case). It returns a p-value, and values below than 0.05 indicate that the two distributions under study are meaningfully different. Very significant Wilcoxon test p-values are obtained for the ovarian, colon, liver, and kidney data sets (3.11e-04, 2.62e-03, 2.10e-05, and 2.33e-08, respectively), indicative of meaningful and important differences between the *nE* distributions in cancer and normal samples, with cancer samples being considerably less stable than their normal counterparts.

In Figure 3.3a the boxplots representing cancer (in green) and normal control (in yellow) *nE* values distributions are shown. Below, the Wilcoxon-rank p-value indicates how significative the medians difference are between cancer (green) and control normal (yellow) distributions.

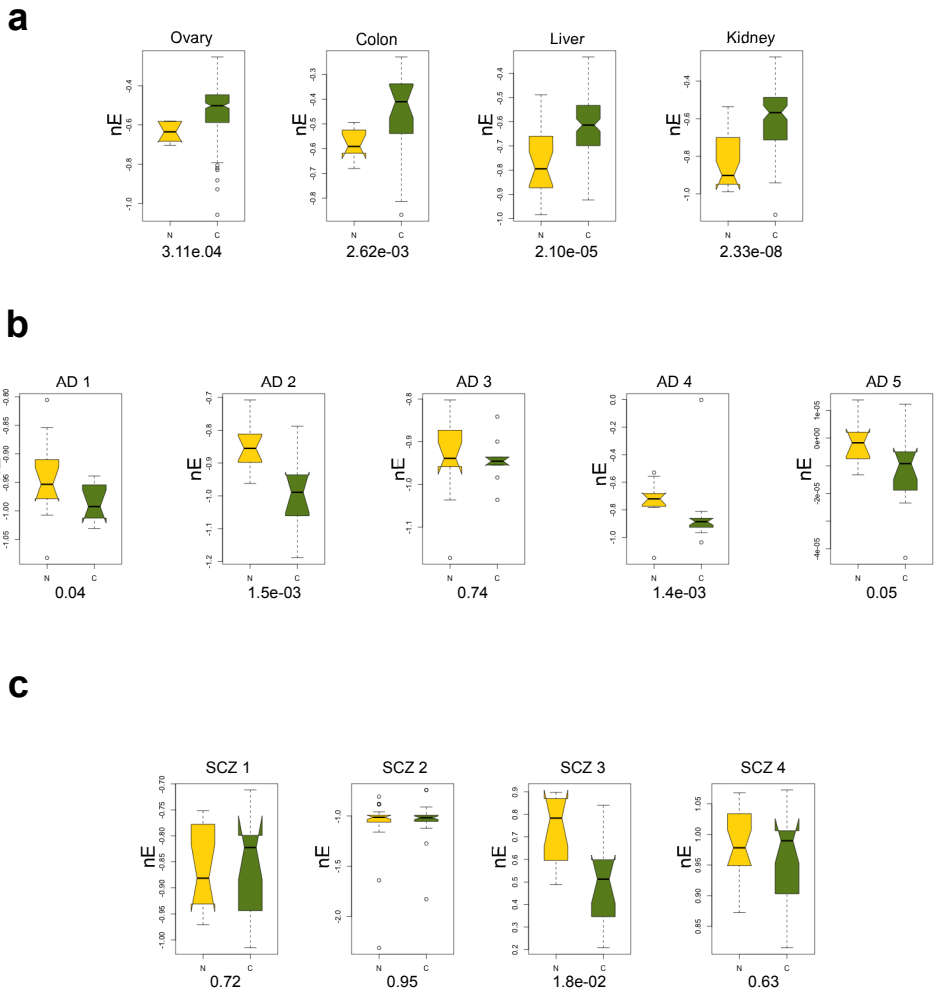


Figure 3.3: The *nE* distribution that maps all the genes in the PINA network in the: (a) normal (N) and cancer (C) states (ovarian, colon, liver and kidney); (b) normal (N) and AD (C); (c) normal (N) and SCZ disease (C) states. The Wilcoxon-rank p-value is presented below the x-axis.

3.5.2 Decreased neighbor-energy in tissues from CNS disorders

Significant differences in *nE* distributions are evident when AD (C) and normal (N) samples are compared (Figure 3.3b), and significant Wilcoxon p-values are obtained for the *nE* distribution in virtually all of the AD studies. AD samples have smaller *nE* values than the normal samples, reflecting increased stability (decreased instability) in the AD network. By contrast, we only observe relevant differences between the *nE* distributions of the normal and disease states for one of the four SCZ data sets available (Figure 3.3c). This discrepancy between the different SCZ networks suggests that further studies are required for this condition, and the underlying cause is unlikely to be revealed until new, high quality experimental datasets become available.

In Figure 3.3b the boxplots representing AD (in green) and normal control (in yellow) *nE* values distributions are shown. Below, the Wilcoxon-rank p-value indicates how significative the medians difference are between AD (green) and control normal (yellow) distributions. In the same way, Figure 3.3c shows the corresponding results for schizophrenia-control cases.

Similar results are obtained when networks other than PINA networks are used, including a smaller HPRD network and a larger HIPPIE one (see Appendix B). It is important to clarify whether these differences are the product of general differences in expression between cancer, normal and neurological disease tissues. However, the normalized expression data (Figure 3.4) indicate that there is no difference between the global levels of normalized expression in this study. This is, Figure 3.4 shows for each disease the distributions of the gene expression data after being normalized. It can be seen, looking at the horizontal line (Q2,median), that there are no significant differences between each case(C)-control(N) gene expression distributions.

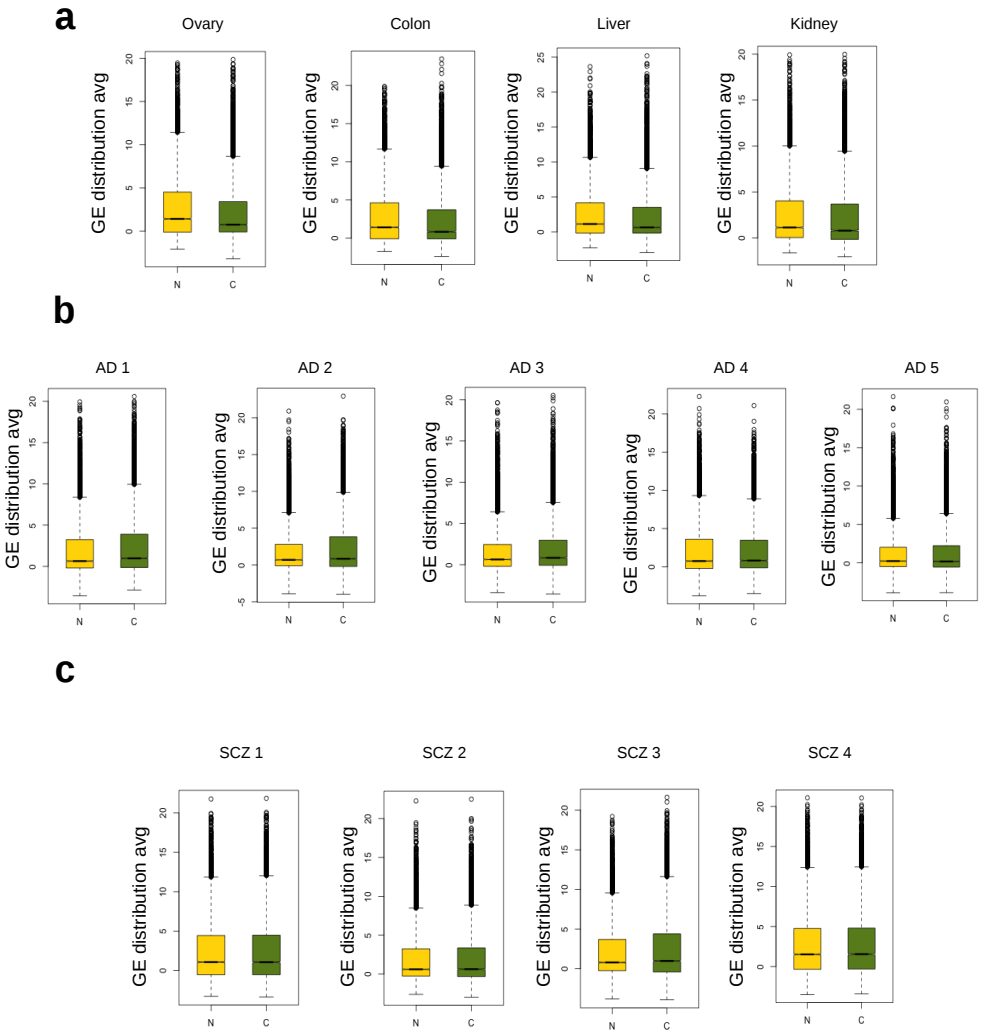


Figure 3.4: Gene expression distribution in: (a) normal (N) and cancer (C) states (ovarian, colon, liver, and kidney); (b) normal (N) and AD (C) conditions; and (c) normal (N) and SCZ (C) conditions.

3.5.3 Consistency of the results

In order to assess the consistency of the results we have analyzed sub-networks obtained by randomly sampling the complete network. Accordingly, 86% of the sub-networks containing 10% of the proteins of the original PINA network produce similar results to the complete network. In other words, not only is there significant instability in the overall network but most of the regions of the network conformed to this behavior, with only a few of them behaving distinctly (Figure 3.5 which includes the *nE* scores for the first one hundred random sub-samples).

For such aim, we have selected aleatory genes from the network, and normalized the overall *nE* score to the total number of genes in the network. We have created 100 sub-networks with different sizes ranged from 1,000 to 1,500 proteins (Figure 3.5), yet similar results are obtained with random sub-sampled networks of 500 or fewer proteins. The result we obtain with the whole network is not trivial but rather, it appears to be quite significant. Figure 3.5 shows 100 *nE* distributions for normal (N, in yellow) and cancer (C, in green) gene expression datasets, for which random sub-networks (selecting 1,000-1,500 random proteins) have been created.

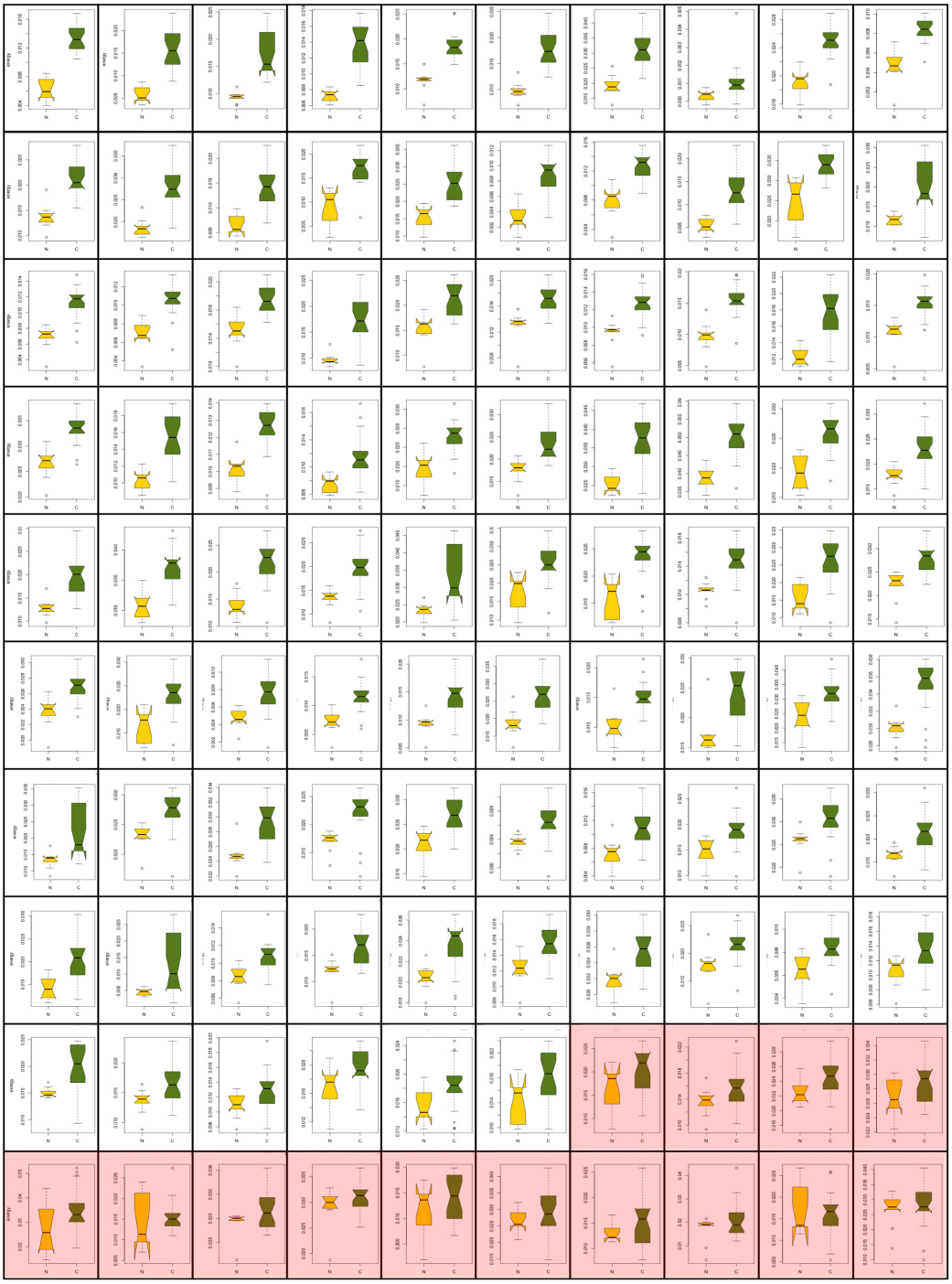


Figure 3.5: The *nE* distribution mapping all the genes in 100 random sub-sample networks in the normal (N) and cancer (C) conditions, sorted by increasing p-values, from left to right and from the top to the bottom. In red, cells with non-significant differences in the *nE* scores between the N and C conditions are shown, representing 14% of the random sub-networks.

3.5.4 Increased neighbor-energy in cancer evolution

To further study the network instability in cancer, we have assessed whether tumor progression might be related with increased instability. For such objective, we have searched gene expression datasets for data associated with different stages or phases of cancer. Using two cancer datasets (colon and liver cancer) that contain gene expression from different tumor stages (see Table 3.1) we have checked whether network instability is affected by the progression of cancer (and so, as time evolves). Figure 3.6 shows the *nE* distributions in different stages of the cancer. In Figure 3.6a it can be seen that the median of the *nE* distribution increases with the level of the liver cancer. The first boxplot corresponds with the normal phase, and the next boxes match with very early hepatocellular carcinoma (HCC), early HCC, cirrhosis with HCC, and advanced HCC (crescent order in disease). A similar outcome is observed in the colon cancer evolution (Figure 3.6b). The median of the *nE* distributions increases while the stages are worse in the cancer development. Hence, the initial results show a significant increase in network instability when the datasets obtained at different stages of tumor progression are compared

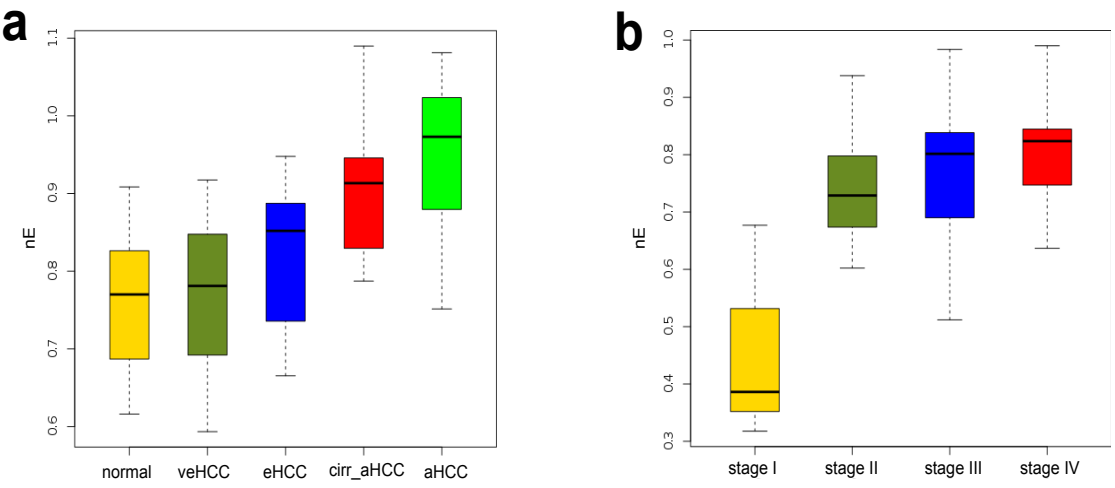


Figure 3.6: The *nE* distributions mapping all the genes at: (a) evolving stages of hepatocellular carcinoma (HCC): normal, very early HCC (veHCC), early HCC (eHCC), cirrhosis with HCC (cirr_aHCC) and advanced HCC (aHCC); and (b) progressive stages of colon cancer stages I, II, III, and IV.

3.5.5 Network stability towards perturbations

Stability has previously been described as a relatively invariant network state when perturbations are introduced. Thus, it is necessary to perform additional experiments to show that our definition of network stability measured through the *nE* score correlates well with this classical definition of robustness. Removing nodes from a network and then studying the evolution of the network's connectivity provides a natural model to study the robustness of networked systems (Callaway et al., 2000; Cohen et al., 2000; Iyer et al., 2013). Changes in the network structure are evaluated in term of the size of the largest connected component of the network. Networks in which the largest component decreases faster than that of the original network are considered to be less robust to perturbations. Thus, nodes have been removed in decreasing order of their *local_nE* scores (Equation 3.2), and accordingly the proteins (or nodes) with higher *local_nE* values (i.e., with more active connections) have been removed first, and proteins with the lowest *local_nE* scores (i.e., less connected with their neighbors) have been removed last. Accordingly, network robustness has been measured by the $R-index$ in Equation 3.4, where α corresponds to the size of the largest connected component within the network after a node is removed. In this manner, the $R-index$ can be used to quantify network robustness (see Section 3.4.5).

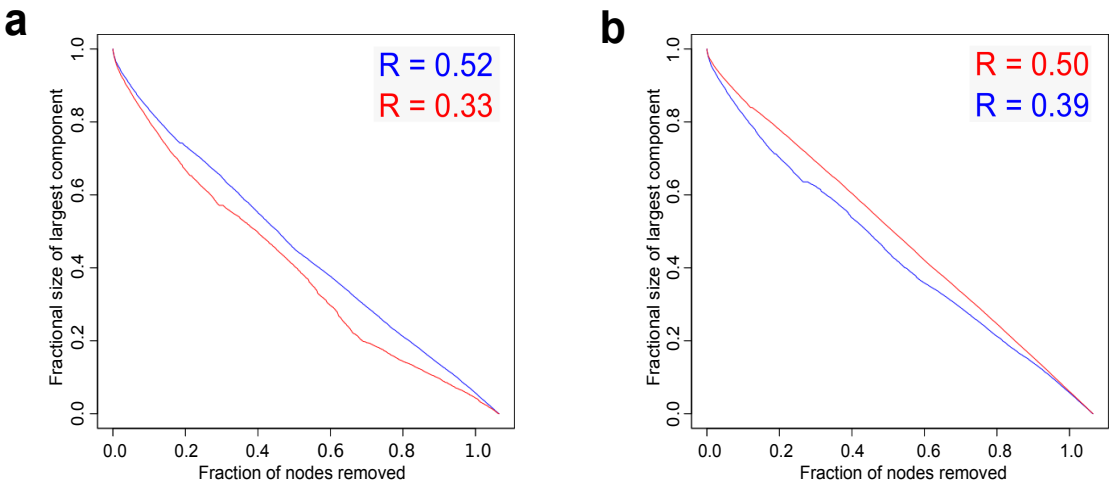


Figure 3.7: Perturbation robustness against *local_nE* sorted score in: (a) cancer samples (red) and normal controls (blue); and in (b) AD samples (red) and normal controls (blue).

The successive removal of nodes according to their *local_nE* score produces a significant difference between the perturbation robustness in cancer and normal samples (Figure 3.7a), and in AD and normal samples (Figure 3.7b). When nodes are removed in a descending order of *local_nE* scores, greater robustness is evident in normal control networks (*R-index* = 0.52) than in cancer networks (*R-index* = 0.33: Figure 3.7a). By contrast, AD networks are more robust (*R-index* = 0.50) than their corresponding normal control networks (*R-index* = 0.39: Figure 3.7b). Hence, the definition of the *nE* score appears to be closely associated to network stability and as such, with the network's robustness to perturbation.

3.5.6 Decreased instability in biological pathways implicated in Alzheimer's disease

We have analyzed the decreased network instability observed for AD samples in more detail and in particular, we have investigated the possible role of the proteins implicated in vesicle trafficking at synapses. Communication between neurons is mediated by the release of neurotransmitter from SVs and the expression of a group of genes involved in SV trafficking is reduced in brain tissues from AD cases. Indeed, the loss of synapses has been correlated with cognitive decline in AD and a malfunction of SV trafficking could be implicated in disrupting neuronal circuits in AD (Yao et al., 2003).

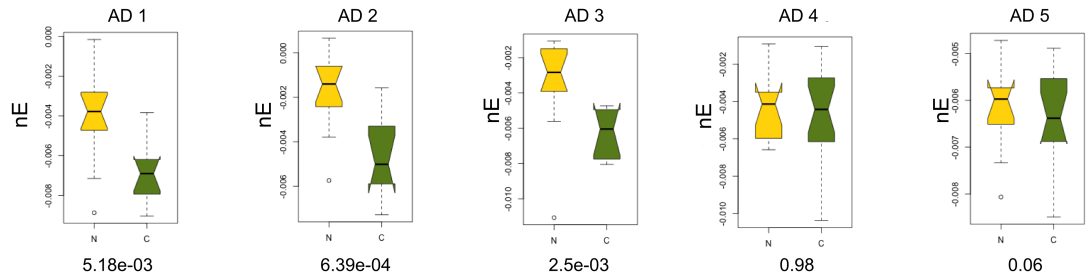


Figure 3.8: The *nE* distribution mapping all the genes in the normal (N) and disease (C) states in AD into the sub-network created from proteins involved in the synaptic vesicle cycle. The Wilcoxon-rank p-value is indicated below the x-axis.

As for the complete PPIN, there is a consistent decrease in instability in the SV related sub-network of proteins from AD samples (Figure 3.8). The difference in the *nE* score suggests that important hubs within the network are expressed and regulated in opposite directions in AD and normal samples. Indeed, nine genes related to endocytosis are expressed in opposite manners in normal and AD samples: *KIT, CLTA, CLTB, AP2M1, AP2S1, AP2B1, HLA-B1, AP2A2,* and *RAB11FIB2*. Three genes associated with SV trafficking (*SYP, STX1A* and *UNC13B*) are inversely expressed in both conditions and they are highly connected in the protein network (hubs). In particular *STX1A* (syntaxin 1A) is known to regulate the exocytosis of SVs and neurotransmitter release (Bennett and Scheller, 1993; Greengard et al., 1993; Hosaka et al., 1999). There is a clear trend towards reduced *STX1A* expression in all AD samples, which has a lower *nE* score than in normal control samples. Indeed, when the *STX1A* gene is not expressed (in blue) nor are its neighbors and conversely, when the *STX1A* gene is expressed (in red) so are most of its neighbors (Figures 3.9 and 3.10). Accordingly, the stability of a particular sub-network relevant to a neurological disease under study is affected in the same way as the stability of the entire network.

The following figures (Figures 3.9 and 3.10) represent PPI sub-networks created from proteins that are involved in the SV cycle in the disease and normal states for AD respectively. Blue nodes represent non-expressed gene products, red nodes expressed gene products, red edges represent interactions between proteins in which both genes are expressed and gray edges represent other combinations. The red clouds contain the STX1A protein as well as all of its interacting partners

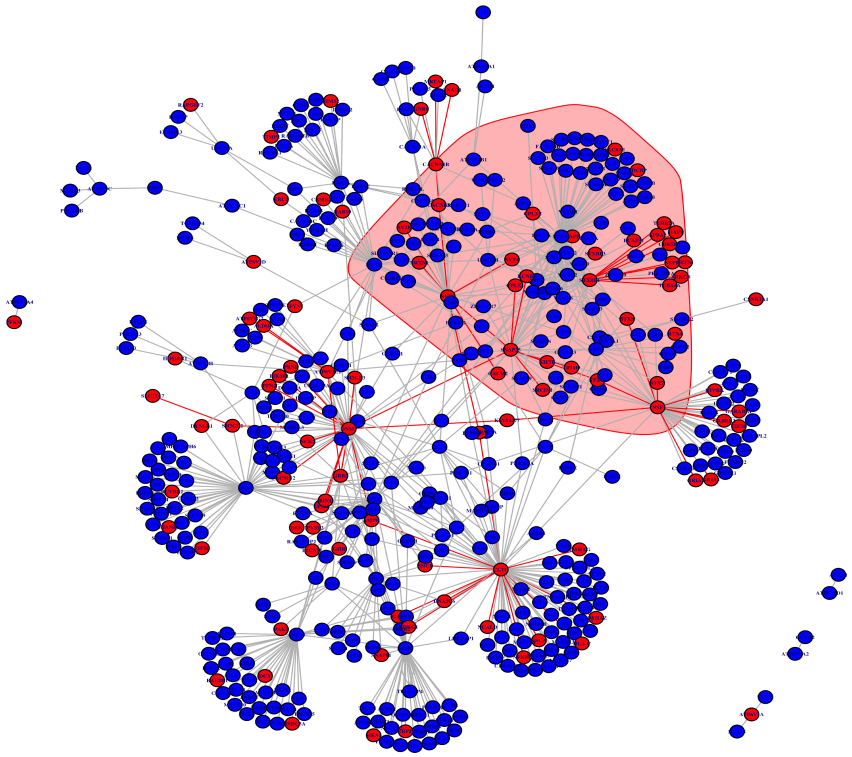


Figure 3.9: Network study of a particular pathway associated with the synaptic vesicle cycle - disease case.

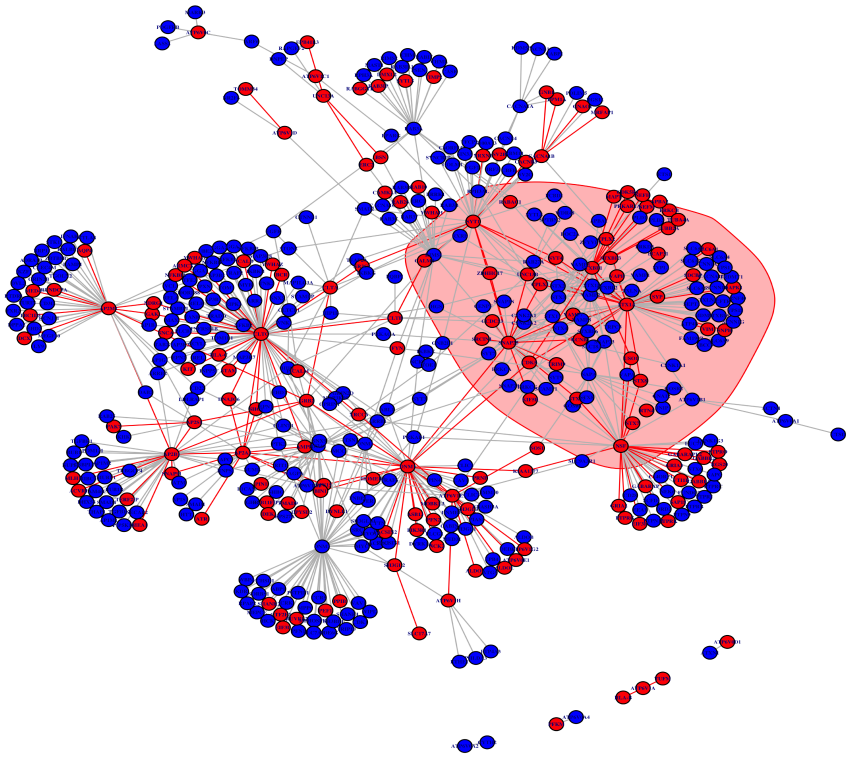


Figure 3.10: Network study of a particular pathway associated with the synaptic vesicle cycle - normal case.

3.6 Discussion

We have designed an approach inspired on DSA, representing PPINs as systems of nodes that are dynamically updated towards a global state of stability. Our strategy is based on the definition of a neighbor-energy function that measures the stability of the network in the general deterministic approach, where *nE* indicates network stability, and it can be interpreted in terms of resistance to alterations and perturbations. In this study, we have analyzed a large set of experimental data on gene expression and various PPINs.

The first significant finding of this study is that networks containing information about expression in four human cancers (ovarian, colon, kidney, and liver) are less stable than the control networks of normal samples. Moreover, this instability in the network seems to increase as these cancers evolve, at least in the tumor progression data sets analyzed. The approach employed is based on the analyses of samples in different conditions and it does not include temporal evolution per se. Thus, the results obtained by analyzing the temporal progression of tumors can be taken as an indication of network evolution towards a less stable state and a way of reconciling our methodology with the standard SA applications.

The randomness or disorder in the local flux distribution surrounding any given node in the network i has been quantified (West et al., 2012), showing that cancer is characterized by an increase in network entropy. This observation could be considered as independent confirmation of our general conclusion. Indeed, when gene expression data is previously integrated with a PPIN for six cancer tissues (Teschendorff and Severini, 2010), an increase in network entropy is again seen to be associated to cancer based on a fluctuation theorem of dynamic systems theory. At the biological level cancer has been associated with a general destabilization of cellular processes related to the organization of the genome, its replication and repair (Murga and Fernández-Capetillo, 2007). A conceptual framework explains how mutations in genes that control genetic stability are selected during tumor progression (Negrini et al., 2010; Loeb, 2011; Wadhwa et al., 2013; Solé et al., 2014). Therefore, our observation of network instability in cancers fits well with current ideas in this field.

Technically, our approach offers important advantages. First, raw gene expression data sets are divergent and independent, which represents an important difference. Additionally, we use a high quality filtered and curated PPIN, which while having practically the same number of total nodes it is less connected than those used in earlier studies. To deal with our biological problem we need to consider both the state of the nodes as well as the strength of the connections between them. This is possible with methods where these two important issues are considered, such as DSA, one of the generic means to resolve the optimization problem (S. Kirkpatrick et al., 1983).

Our second important finding is that the AD network is more stable than its control normal network, with a significant increase in the *nE* of the corresponding networks. This is an interesting behavior that contrasts with that of cancers, and as far as we know is detected here for the first time. One possible interpretation of these results would be that cancer implies a general deregulation of cell growth through the hyperactivation of certain pathways, resulting in a destabilization of their interactions, while AD and other neurological disorders imply the stabilization of biological processes and network interactions, and their general slowing down. The striking contrast in the behavior of cancer and AD networks, from less to more stable networks, should be considered in the context of the observed *inverse comorbidity* of these two groups of diseases. A substantial number of epidemiological studies have shown that there is an inverse relationship between cancer and several central nervous system diseases, including AD. In other words, patients with AD tend to less frequently suffer some types of cancer (Tabarés-Seisdedos et al., 2011; Tabarés-seisdedos and Rubenstein, 2013) [for a complete meta-study of the available epidemiological studies see (Catalá-López et al., 2013)].

Finally, given the importance of the diseases discussed in this work, it is necessary to make these results accessible for future experimental analysis. In this sense, an initial study of the molecular basis of this *inverse comorbidity* identified sets of genes expressed weakly in AD and strongly in cancers (Ibáñez et al., 2014). The new methodological approach developed here represents a further advance with respect to that initial approximation, where genes are not considered as independent units but rather as part of a connected network. This approach could be used as a classifier to distinguish cancer and normal samples. Another possibility will be to cluster the results of this procedure in order to extract specific proteins for which additional experimental information could be available, or could be tracked in direct experiments.

Furthermore, this scheme could be also applied to any network system where the elements are characterized by a state S_i and their interactions associated to a weight W_{ij}. In a biological context there are numerous systems with these characteristics, such as protein interaction and gene control networks. In the future, the application of clustering techniques to disease networks, such as self organizing maps (SOM), will render information not on single genes but on clusters of collaborating genes, moving towards the study of the molecular causes of comorbidity to the level of systems biology.

4 Recognition between cancer and CNS disorders related proteins

4.1 Summary

A pattern recognition system is presented, in which proteins related with CNS disorders and cancer related proteins are classified. We have defined two attributes that are capable to categorize in different groups the majority of the proteins related to cancer or CNS disorders, corresponding with clusters with high or low feature values. The *dex* feature is indicative of the number of neighbors of each protein being expressed, important to estimate how connected and active proteins are within the PPIN. We have adopted the *local_nE* attribute from the work presented in chapter 3 and published in (2015), which measures the stability of a single protein in its neighborhood.

4.2 Introduction

The development of large-scale genomics methods and large biological databases in combination with different statistical approaches and adequate machine learning techniques build the basis for the analysis of the relations between human complex disorders. For instance, integrating multiple 'omics' analyses in combination with a genetic algorithm (Liu et al., 2013b) identifies essential biomarkers in preeclampsia, a bioinformatics-based drug approach is capable of repurposing already approved drugs and identifying novel class of molecules to treat diseases (Jahchan et al., 2013) or combining gene expression and single nucleotide polymorphisms successes recovering, and predicting known therapeutic targets across several human diseases through a systematic approach (Fan-Minogue et al., 2015).

Gene expression determines whether the gene is expressed or not. Most genes code for proteins that are the ones carrying out most biological processes. Many protein functions take place in the context of the interactions with other proteins, which are described in the PPIN, in which each node represents a protein and the edges the interactions with other proteins. It is not possible to obtain a network for each condition (disease or normal), and thus, an alternative is to combine gene expression data with PPIN to gain a better understanding of the organization and level of activity of the network (Srihari et al., 2014; de Baumont et al., 2015; Chen et al., 2015; Xiao et al., 2015). The strategy of combining PPINs with gene expression is artificial, since the networks that are available correspond to a particular state rather than diseases, individuals or conditions. Numerous studies have used PPIs and gene expression information to study complex disorders (Chuang et al., 2007; Pujana et al., 2007; Hudson et al., 2009; Milanesi et al., 2009; Komurov et al., 2010; Schramm et al., 2010; Teschendorff and Severini, 2010; West et al., 2012; van Pel et al., 2013), and investigated the properties of the corresponding networks (Schadt, 2009).

The recent development of more comprehensive and accurate PPINs provides additional possibilities in this area (Cusick et al., 2009; Barabási et al., 2011; Kerrien et al., 2012; Licata et al., 2012; Chatr-Aryamontri et al., 2013; Rolland et al., 2014). Protein interaction maps have been used to study the molecular organization of cellular systems and the perturbations created by disease (Dobson et al., 2014; Safari-Alighiarloo et al., 2014). In particular, topological analyses of PPINs show that cancer proteins tend to be central within the network as they are highly connected (Jonsson and Bates, 2006; Taylor et al., 2009; Sun and Zhao, 2010; Xia et al., 2011; Choura and Rebaï, 2012; Xiong et al., 2014). It has been proposed that important proteins for a cell's survival are highly connected (Jeong et al., 2001) and altering them has profound effects on the interaction network. Similarly, studies on the genomic and network characteristics of genes that mutated in cancer indicate that these genes tend to encode central hubs within a PPIN (Rambaldi et al., 2008). Correspondingly, Xia et al. (2011) argued that cancer-related proteins (CRPs) have a much stronger protein–protein interaction density than control proteins in the whole human interactome. The authors retrieved cancer and non-cancer essential genes from different repositories and studied the different topological features within the PPIN, such as the degree, betweenness, and centrality measures. In this study, we examine critically this hypothesis by comparing the network organization of proteins directly related with two different disease groups

(cancer related proteins CRPs and proteins related with neurological disorders - NRPs). It is particularly interesting to examine the two groups of diseases given the active debate on the medical and biological relations between cancer and neurological or CNS disorders.

At the medical level, population and epidemiological studies have shown that some types of cancers and CNS disorders co-occur less often than expected. Indeed, patients with certain neurological disorders, including those suffering from Parkinson's or Alzheimer's disease, have a lower than expected tendency to develop some forms of cancer (Behrens et al., 2009; Tabarés-Seisdedos et al., 2011; Behrens et al., 2012; Tabarés-Seisdedos and Rubenstein, 2013). A phenomenon called *inverse comorbidity*.

At the molecular level, we detected a common set of genes and biological pathways which expression levels are regulated in opposite directions in CNS disorders and cancer [described in chapter 2 and published in (Ibáñez et al., 2014)]. Later using a simulated annealing inspired approach, we showed that CNS disorders are characterized by lower network instability while networks informed with cancer gene expression data tend to have higher instability (Ibáñez et al., 2015). Our interpretation is that the larger number of interactions among CRPs makes the network more susceptible to destabilization. Similarly, West et al. (2012) and Teschendorff et al. (2010) found that cancer cases have an increased network entropy using a different computational approach. This study aims to clarify the origin of the differential network behavior of CRP and NRPs, by studying their network neighborhoods with different clustering methodologies. Clustering is an elemental problem with many applications in biology, medical research, bioinformatics, and other disciplines. Particularly, self-organizing maps (SOM) have been used to recognize and classify features in human hematopoietic gene expression data. A hierarchical clustering in the budding yeast *Saccharomyces cerevisiae* was proposed (Eisen et al., 1998) in which clustering gene expression data efficiently grouped together genes of known similar function. Graph theoretic approaches, among others, were also used to group together genes of known similar function or genes with particular features (Yeung et al., 2001; Ben-Dor et al., 2004).

4.3 Material

4.3.1 Microarray gene expression data

Gene expression experimental raw data (CEL files) have been downloaded from the Barcode human transcriptome repository (bar) for ovarian, colon, and liver cancer samples. For the CNS disorders, diseased samples have been downloaded from the NCBI GEO omnibus (geo) and the Stanley Medical Research Institute Online Genomics Database (smr) for schizophrenia (SCZ), Alzheimer's disease (AD), and Parkinson's disease (PD) (see Table 4.1).

Each CEL file includes probe intensities information produced at the end of the microarray scan, and this data must be preprocessed (see Section 4.4.1). Table 4.1 incorporates the experimental gene expression datasets used. The tissue field indicates the human region in which the experiment is done, the platform type the official name of the microarray platform used, the sample size the total number of patients included (control and cases), and the source the identifier of the gene expression dataset in the original repository.

Table 4.1: Gene expression datasets.

Tissue	Platform	Sample Size	Source
Alzheimer's disease			
Entorhinal Cortex	HG-U133Plus2	23	GSE5281
Hippocampus	HG-U133Plus2	23	GSE5281
Medial Temp. Gyrus	HG-U133Plus2	28	GSE5281
Posterior Singulate	HG-U133Plus2	22	GSE5281
Primary Visual Cortex	HG-U133Plus2	31	GSE5281
Superior Frontal Gyrus	HG-U133Plus2	34	GSE5281
Hippocampus	HG-U133Plus2	24	GSE1297
Schizophrenia			
Postmortem cerebellum	HG-U133Plus2	28	GSE4036
Postmortem frontalBA46 cortice	HG-U133Plus2	55	Dobrin
Postmortem hippocampus (CA1)	HG-U133Plus2	41	Laeng

Table 4.1 – Gene Expression datasets

Tissue	Platform	Sample Size	Source
Postmortem thalamus (MD)	HG-U133Plus2	26	Kemether
Parkinson's disease			
Postmortem thalamus	HG-U133Plus2	26	GSE7621
Postmortem medial, lateral and frontal	HG-U133A	45	GSE8397
Substantia nigra	HG-U133A	29	GSE20292
Colorectal cancer			
Colon tissue	HG-U133Plus2	33	GSE4183 GSE7307 (GSM175905) GSE2109
Ovarian cancer			
Ovary tissue	HG-U133Plus2	278	GSE9891 GSE2109 GSE9890 GSE3526 GSE7307
Liver cancer			
Liver tissue	HG-U133Plus2	129	GSE6222 GSE9829 GSE9843 GSE6956 GSE11045 GSE7307

4.3.2 The protein–protein interaction network

We have used the human protein interaction network from the Protein Interaction Network Analysis (PINA) database. PINA (pin) is an integrated platform of PPI data extracted from six different public databases: IntAct, MINT, BioGRID, DIP, HPRD, and MIPS/MPact. It includes self-interactions, interactions predicted by computational methods, and interactions between human proteins and proteins from other species. For our purpose, we first have used data from the PINA website (pin) and then filtered the data by requiring PPIs to have experimental evidence and removing redundancy and self-interactions as well as interactions involving proteins not from *Homo sapiens*. We only have considered the interactions among proteins that were also detected in the microarray platforms (Figure 4.2b). The resulting filtered PINA network consists of 10,650 proteins with 63,119 interactions. Each node denotes a protein encoded by a gene, and each edge denotes an interaction existing between two proteins. PINA has also been used in recent studies, including those of Xia et al. (2011) and Laakso et al. (2010) among others.

It is important to emphasize the dimension and complexity of the data. For each disease, different experimental gene expression datasets are retrieved (biological replicates) with cases and control samples. Each sample includes 10,650 genes. Accordingly, we are working with 875 samples containing each of one 10,650 genes.

4.4 Methods

We have designed a pattern recognition system using a SOM clustering methodology, and we have validated further with K-Means. The corresponding flow chart is presented in Figure 4.1. In the first step, preprocessing each individual microarray is necessary to estimate the expression level of each gene on the array (Section 4.4.1). Defining the features that the clustering method classifies into different kinds of groups or clusters is fundamental (Section 4.4.2). Thus, we have defined *dex* and a *local_nE* (defined below), this last attribute is based on our previous study (Ibáñez et al., 2015), as the principal characteristics of each gene to be classified. To do this, we have normalized all the data (Section 4.4.3), and analyzed the optimal number of clusters (Section 4.4.4) in which the genes are more effectively grouped (Section 4.4.5).

A SOM clustering algorithm is developed on the space of *dex* and *local_nE* for every gene in the cancer and neurological samples (Section 4.4.4). This step is validated with the K-Means approach. Finally, we examine each cluster independently by analyzing which groups with higher *local_nE* and *dex* values are more enriched in the CRPs than in the NRPs, and vice versa.

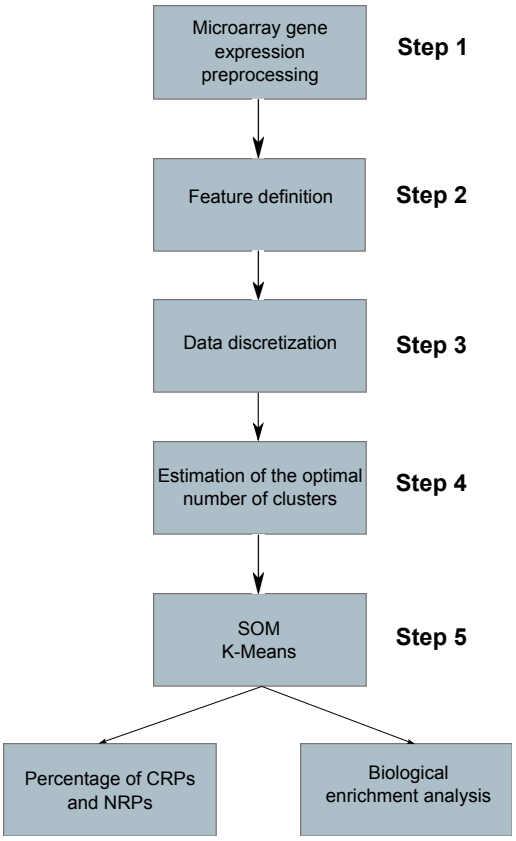


Figure 4.1: Flow chart of our pattern recognition system.

4.4.1 Microarray gene expression preprocessing (Step 1)

The preprocessing of each single microarray sample is an essential step to estimate the expression of each gene on the array. The SCZ, AD, and PD samples in the CNS disorders and the ovarian, colon, and liver cancer samples are normalized using frozen Robust Multiarray Analysis (fRMA) (McCall et al., 2012) from the R affy package (Gautier et al., 2004). The fRMA processes each array individually and accounts for probe variability, batch effects, probe effects, array-to-array variability, and background noise, obtaining background-corrected gene-level intensities. Afterwards, the methodology proposed by (Zilliox and Irizarry, 2007) have been used to map the gene intensities (Z-score) into a vector of ones and zeros that denote which genes are expressed (ones) and unexpressed (zeros) for each sample (Figure 4.2a). To compare the Z-score among all the diseases, we have processed it following the strategy described at section 3.4.3.

In summary, two different data types are obtained for each sample: the significance level (Z-Score) of the gene being expressed or not, and the information on whether the gene is expressed (one) or not (zero) (Figure 4.2a).

a Gene expression

Gene	Z-score	Normalized Z-score (S)	Barcode
DDR1	8.24	1	1
RFC2	1.49	0.07	0
PAX8	6.79	0.97	1
GUCA1A	-0.66	0.6	0
UBA7	6.86	0.98	1
THRA	-0.01	0.5	0
PTPN21	0.43	0.33	0
CCL5	5.02	0.99	1
BRCA1	-0.23	0.40	0
BRCA2	4.56	2.5e-06	0
RAD51	0.19	0.42	0
BARD1	3.21	6.63e-04	0
RBBP8	3.84	6.15e-05	0
MAPK1	4.65	1.65e-06	0
VPS18	0.24	0.4	0
.	.	.	.
.	.	.	.
.	.	.	.
PCP2	-1.26	1	0
MTCP1	-0.06	0.62	0
NAF1	2.78	2.7e-03	0
ATP6V0C	12.10	1	1

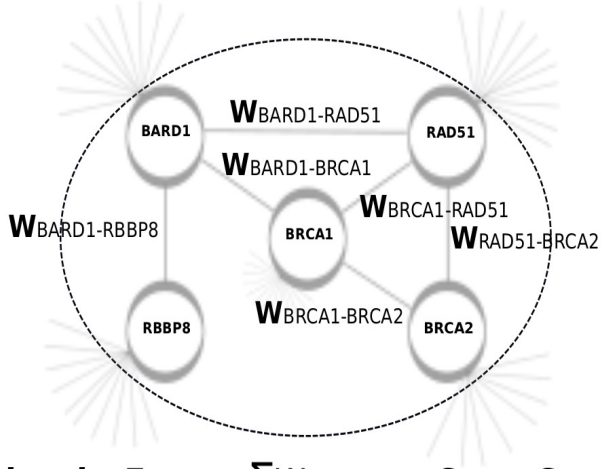
b Protein interactions

GeneA	GeneB
BRCA1	BRCA2
BRCA2	RAD51
RAD51	BRCA1
RAD51	BARD1
BARD1	BRCA1
RBBP8	BARD1
BRCA1	RBBP8
CHEK2	STRAP
GH1	BTRC
WNK1	MAP3K2
CLTC	STAMBPL1
MLLT10	DOT1L
ESR1	CUL4B
FLOT1	ARG1
NFKB2	DYNC1H1
.	.
.	.
.	.
FOS	RNF20
ERBB2	NRG1
MYC	EP400

c Feature assignment

Gene	dex	local_nE
DDR1	8	10.58
RFC2	4	0.05
PAX8	6	8.62
GUCA1A	0	-0.47
UBA7	1	0
THRA	15	2.19
PTPN21	0	-0.67
CCL5	4	-5,62
BRCA1	62	15.51
BRCA2	14	0.95
RAD51	13	0.02
BARD1	13	0.21
RBBP8	6	12.05
MAPK1	52	2.05
VPS18	3	0.13
.	.	.
.	.	.
.	.	.
PCP2	5	0.59
MTCP1	1	-0.01
NAF1	1	-0.01
ATP6V0C	2	2.06

d Example subnetwork



$$local\_nE_{(BRCA1)} = -\sum W_{BARD1\text{-}BRCA1} * S_{(BARD1)} * S_{(BRCA1)},$$
$$W_{RAD51\text{-}BRCA1} * S_{(RAD51)} * S_{(BRCA1)},$$
$$W_{BRCA1\text{-}BRCA2} * S_{(BRCA1)} * S_{(BRCA2)},$$
...

Figure 4.2: (a) Subset of the preprocessed and normalized gene expression data. (b) Subset of the filtered PPIN. (c) Subset of the *dex* and *local_nE* values for each gene. (d) Example of the BRCA1 *local_nE* computation in a subnetwork.

4.4.2 Feature selection (Step 2)

The selection of features is essential when a classification algorithm is used. The number of features involved in the analysis directly affects the performance and robustness of the methodology (Jain and Chandrasekaran, 1982).

For each protein in the PPIN, the *dex* and *local_nE* attributes are defined. The *dex* feature indicates the degree of each protein "expressed" in the PPIN, that is, the number of neighbors of each protein being expressed. This feature is important to estimate how connected and active a protein is within the network. The degree of proteins is an invariable attribute in the PPIN. For the *local_nE* attribute, we adopt the *nE* function (Ibáñez et al., 2015) described in chapter 3. The original *nE* function measures the stability of an entire network. Stability describes a network state that is not significantly altered when its properties or the perturbations within it are changed. In this new study, we take advantage of the *local_nE* function, which measures the stability of a single protein in its neighborhood, that is, only with the direct interacting partners, and not in the entire network as in the original work (Ibáñez et al., 2015). Our principal objective is to determine the stability of each protein with its neighbors, rather than within the whole network. To make our goal more precise, the definition of *local_nE* is fundamental to determine the effect and alterations a protein can produce in its neighborhood.

The system is represented by a PPIN, and nodes (S_i) represent the proteins associated with the value of expression of the corresponding gene. Each S_i describes the significance level of the gene i being expressed or not. Edges (W_{ij}) show the existing interactions among proteins, and are defined in Equation 4.1. This equation is equivalent to 3.1, but it is shown here again for better reading.

$$W_{ij} = \begin{cases} -1 & \text{if i and j are expressed} \\ +1 & \text{if i or j are not expressed} \\ +1 & \text{if i and j are not expressed} \end{cases} \tag{4.1}$$

Following the main idea of the deterministic simulated annealing algorithm, the *local_nE* function defined (Ibáñez et al., 2015) is the sum of the energy of all nodes connected to a given node i. These influences are calculated by multiplying the level of expression of each node (S_i) by the associated weights, with all the nodes interconnected with the first one (W_{ij}), as expressed in Equation 4.2. This equation is equivalent to 3.2, but it is shown again here for better reading.

$$local\_nE(i) = -\sum_{j} W_{ij} * S_i * S_j \tag{4.2}$$

According to the definition in Equation 4.2, *local_nE* is maximal when $W_{ij} * S_i * S_j$ is at its minimum. It represents the active connections among the nodes of the expressed genes (Equation 4.1, case 1) and indicates that any alteration in this node will destabilize the network. The value of *local_nE* decreases for node connections that involve at least one gene that is not expressed in that condition, thus indicating the fact that the interactions cannot take place (Equation 4.1, cases 2 and 3). In this situation, the *local_nE* achieves its minimum value, thus indicating network stability.

Consequently, the *local_nE* function measures the stability of a single protein or node in the function of its neighborhood, that is, only with the direct interacting partners and not in the entire network.

4.4.3 Data discretization (Step 3)

The SOM and K-Means algorithms we adapt here use the Euclidean distance to measure the distance between a data element (protein) and its cluster's centroid. In using the Euclidean distance, the clustering results can be greatly affected by the difference in scale of the dimension from which the distances are computed. Therefore, as these distances are computed from processed raw data, they are normalized to prevent dependence on the choice of measurement units (Han et al., 2011). The possibility that scores with the largest range could dominate the distance computation is avoided. In this manner, *local_nE* and *dex* features are normalized so that each feature has mean 0 and variance 1. For such aim, the $normalize$ function of the som R package (som) is used. This function normalizes the data so that each column has

mean 0 and variance 1 (each column represents each feature).

4.4.4 Determination of the optimum number of clusters (Step 4)

The number of clusters is initially unknown. No predefined classes exist before grouping, and finding an appropriate metric for measuring whether the found cluster configuration is acceptable or not is a difficult task. Therefore, an evaluation of the clustering methodology is necessary (Legány et al., 2006). Several validity indices are here tested, and similar results are obtained using Calinski-Harabasz (CH_{index}) (Calinski and Harabasz, 1974) and Davies-Bouldin (DB_{index}) (Davies and Bouldin, 1979) validity indices. The R fpc package (fpc) and R clusterSim package (clu) are used for the two validity indices, respectively, to test the outcome of each index for different numbers of clusters.

The idea behind the CH measure (Calinski and Harabasz, 1974) is to compute the sum of the distances between the k-th cluster and the other $k-1$ clusters, and to compare this sum with the internal sum of the distances for the k clusters. This measure is one of inter-cluster (dis)similarity over intra-cluster (dis)similarity. It works as shown in equation 4.3:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/N-k} \tag{4.3}$$

where $B(k)$ is the distance among clusters, $W(k)$ is the distance in the cluster, k is the index of the cluster, and N is the total amount of clusters.

A good clustering is associated with the highest CH value, and it occurs when the difference among the clusters is high, and the within cluster difference is low. The main goal of this step is to predict the optimal number of groups in which data can be organized according to the defined attributes. Accordingly, CH_{index} is used for measuring the "goodness" of the clustering without the need for a manual exploration step (a similar result is obtained with DB_{index}).

4.4.5 SOM and K-Means clustering (Step 5)

In the current study, SOM and K-Means clustering algorithms are developed on the space of *dex* and *local_nE* attributes for every gene in the samples. Clustering solutions are gathered for a pre-specified number of groups, and the solution maximizing CH_{index} (Calinski and Harabasz, 1974) is selected.

A clustering problem consists of elements and a feature vector for each element. A measure of similarity is defined between pairs of such vectors. In our study, the elements correspond with the encoded proteins, and the vector of each protein contains *local_nE* and *dex* values. The main goal is to partition the elements into subsets (i.e., clusters) to satisfy the two criteria defined in (Shamir and Sharan, 2002): homogeneity (elements in the same cluster are highly similar to each other), and separation (elements from different clusters have low similarity to each other).

These classical clustering algorithms, which assume an already known number of clusters, have been widely used in several works (Thalamuthu et al., 2006; de Castro Leão et al., 2009; Omer et al., 2014). The main goal is to minimize the distance among elements and the centroids of their assigned clusters (Shamir and Sharan, 2002). Let M be the $n * m$ matrix. For a partition P, the elements in $\{1,...,\}$ are denoted by $P(i)$, which is the cluster assigned to i, and by $c(j)$, which is the centroid of cluster j. Let $d(v_1, v_2)$ denote the Euclidean distance between the vectors v_1 and v_2. K-Means attempts to find a partition P in which the error function $E_p = \sum_{i=1}^{N} d(i, c(P(i)))$ is minimum.

The current partition is modified by checking all possible modifications of the solution, in which one element is moved to another cluster, and by making a switch that reduces the error function. SOMbrero R package (Olteanu et al., 2012; Bendhaiba et al., 2015; Olteanu and Villa-Vialaneix, 2015) and stats R package (sta) are used for SOM and K-Means clustering respectively.

4.5 Results

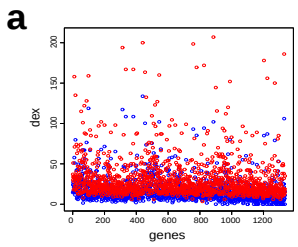
The pipeline proposed in section 4.4 is followed (Figure 4.1) for result analysis and interpretation. Microarray raw data are preprocessed in the first term; the features for the clustering are defined; the optimal number of clusters is determined; both the SOM and K-Means clustering algorithms are conducted; and the cluster results are biologically interpreted.

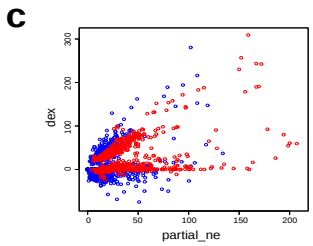
The SOM and K-Means clustering algorithms are applied on the space of *dex* and *local_nE* features for every gene in the cancer and neurological disorder samples with similar outcomes. The optimal number of groups is defined by maximizing CH_{index} (Calinski and Harabasz, 1974) (a similar outcome is obtained by minimizing DB_{index}), as described in section 4.4.4.

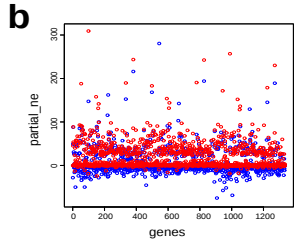
4.5.1 Outcome of the feature selection

Considering *dex* and *local_nE* features independently, we do not observe differentiated groups when these two feature values in the genes expressed in cancer (in red) and CNS disorders (in blue) are represented (Figure 4.3a and 4.3b). However, when both *dex* and *local_nE* values are combined in genes expressed in cancer (in red) and CNS disorders (in blue), two groups are distinguished with differentiate patterns (Figure 4.3c).

To estimate how adequate these defined features are, we have used different classifiers such as Weka (Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, 2009) for Random Tree Forest, J48, and Bayes classifiers with and without stratification. Consequently, genes from the cancer data are labeled “cancer” and those from CNS disorders are labeled “neuro.” Considering *dex* and *local_nE* values the previous classifiers are used to correlate genes according to each feature value to each class ("cancer" or "neuro"). The percentages of genes correctly matched are 98.3%, 85. 7%, and 83.3% using Random Tree Forest, J48, and Bayes classifiers, respectively (Figure 4.3d).







d

	Correctly classified instances	Incorrectly classified instances	Percentage Correctly Classified
Random Forest	1524	26	98,32%
J48	1329	221	85,74%
Bayes	1292	258	83,35%

Figure 4.3: Outcome of the feature selection (a) Representation of the *dex* value in genes expressed in cancer (red) and CNS disorders (blue). (b) Representation of the *local_nE* value in genes expressed in cancer (red) and CNS disorders (blue). (c) Representation of both *dex* and *local_nE* values after normalization in genes expressed in cancer (red) and CNS disorders (blue). (d) Outcome of the Random Tree Forest, J48, and Bayes classifier algorithms considering both *dex* and *local_nE* features.

4.5.2 Optimal number of clusters

The main goal here is to find out the most appropriate configuration of SOM and K-Means in order to represent the topology of the data (see Section 4.4.4). Thirteen K-Means and SOMs of different dimensions with the same data set are trained and clustered, and validation indexes are computed for each configuration. The maximal CH_{index}, reflecting the "goodness" of the clustering, corresponds with 10 (Figure 4.4) (similar result is obtained with DB_{index}). Consequently the K-Means and SOM are going to be divided into 10 different groups (matrix with dimension 2x5).

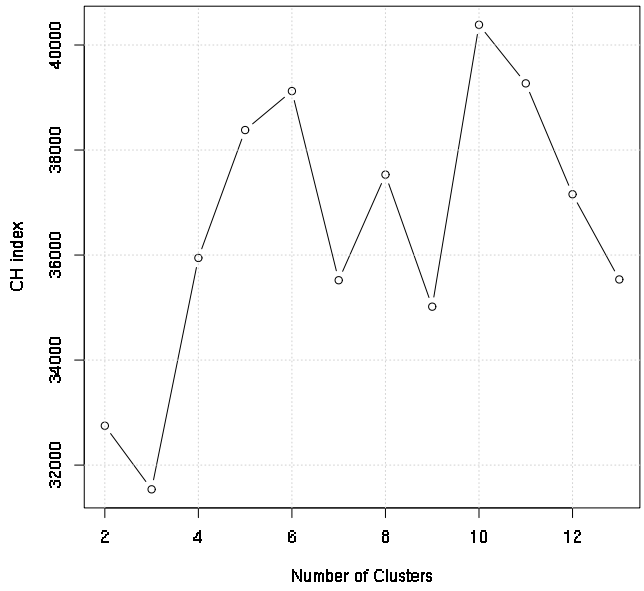


Figure 4.4: Estimation of the optimal number of clusters maximizing CH_{index}.

4.5.3 SOM and K-Means approaches

Following the iterative process of the proposed pipeline, once the optimal number of groups is estimated, both the SOM and K-Means clustering algorithms are performed with similar outcomes. Proteins with a similar number of expressed neighbors and having an approximate *local_nE* are sorted in the same cluster.

The number of proteins associated with cancer and CNS disorders is the same; that is, we analyze the same proteins in all the diseases. If the clusters were generated randomly, the elements inside each group would be adjusted as a binomial distribution, $Bin(n, 0.5)$, where n is the cluster size, and 0.5 is the probability of having a cancer or neurological protein in a group. The binomial p-value is computed based on the number of CRP or NRPs compared to the total amount of proteins within each group. Accordingly, low binomial scores (p-value < 0.01) indicate that the proteins within the cluster do not follow a binomial distribution, and so, the presence of CRP or NRPs is significant.

Clearly separated groups are represented by different colors in SOM and K-Means methodologies (Figures 4.5a and 4.6a). In Figures 4.5a and 4.6a are represented the

proteins by their corresponding *dex* and *local_nE* values after the SOM and K-Means respectively. In Figures 4.5b and 4.6b, each fraction in the pie chart symbolizes the number of proteins related to each disease (AD, SCZ, PD, ovarian, liver, or colon cancer) with the feature values of the corresponding cluster, the p-value score associated with the binomial distribution, and the total number of proteins in each cluster (n).

In Tables 4.2 and 4.3 the percentages of CRP and NRP for each cluster are represented.

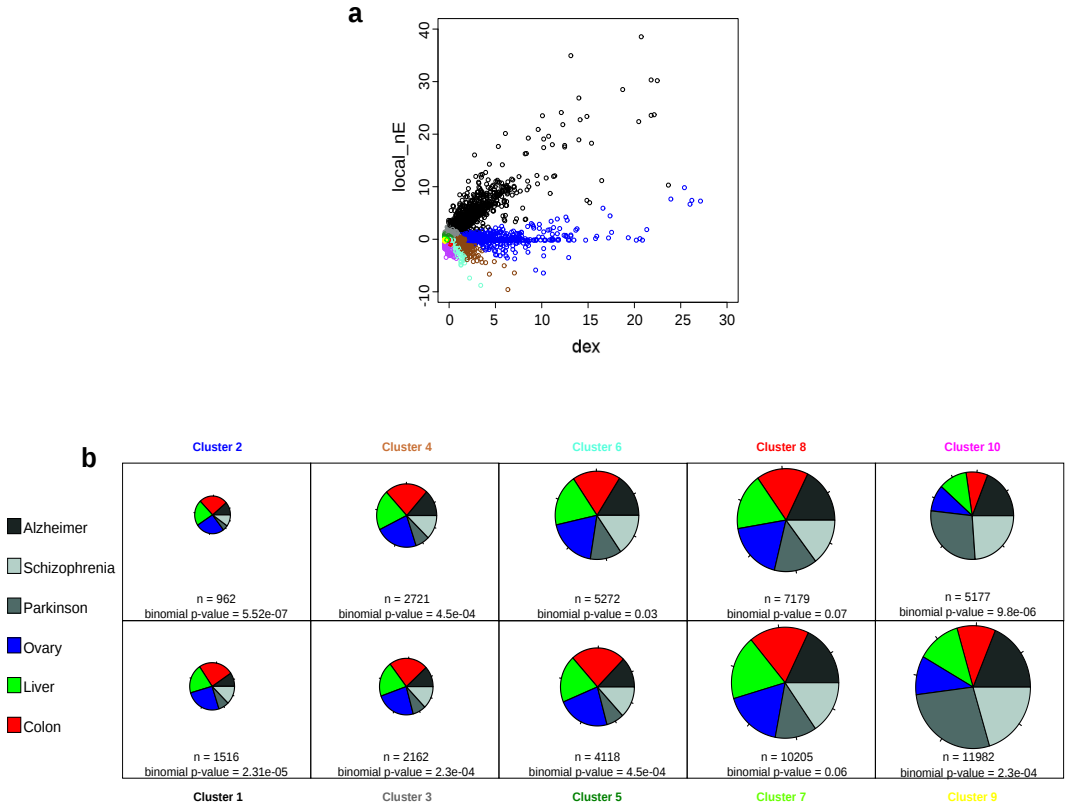


Figure 4.5: SOM clustering analysis (a) Proteins by their corresponding *dex* and *local_nE* scores after SOM clusterization. (b) Pie charts containing the colored fraction of diseases in each cluster characterized by proteins with similar *local_nE* and *dex* values.

Table 4.2: Percentage of proteins in each cluster associated with CNS disorders and cancer with SOM.

Cluster	NRP percentage (%)	CRP percentage (%)
Cluster 1	29.5	70.5
Cluster 2	25.7	74.3
Cluster 3	32.6	67.4
Cluster 4	33.4	66.6
Cluster 5	33.4	66.6
Cluster 6	43.5	56.5
Cluster 7	54.2	45.8
Cluster 8	46.5	53.5
Cluster 9	66.6	33.3
Cluster 10	70.7	29.3

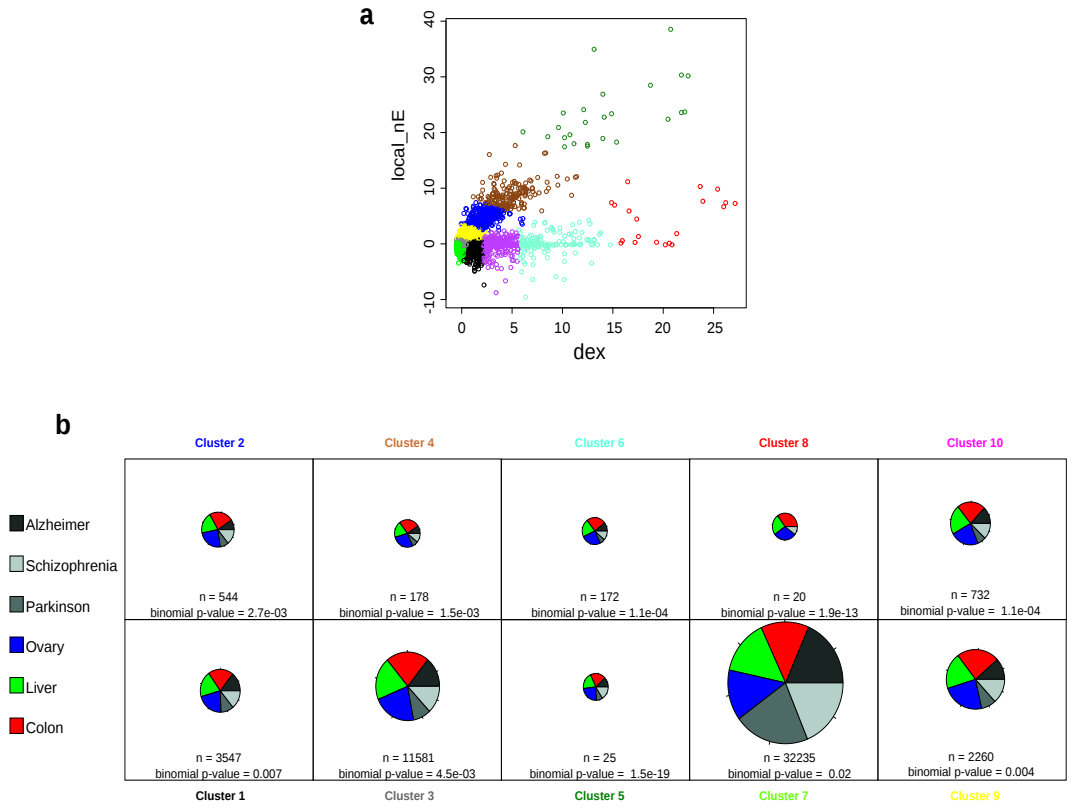


Figure 4.6: K-Means clustering analysis (a) Proteins by their corresponding *dex* and *local_nE* scores after K-Means clusterization. (b) Pie charts containing the colored fraction of diseases in each cluster characterized by proteins with similar *local_nE* and *dex* values.

Table 4.3: Percentage of proteins in each cluster associated with CNS disorders and cancer with K-Means.

Cluster	NRP percentage (%)	CRP percentage (%)
Cluster 1	38.7	61.3
Cluster 2	31.9	68.1
Cluster 3	36.2	63.8
Cluster 4	27.5	72.5
Cluster 5	36.0	64.0
Cluster 6	27.9	72.1
Cluster 7	63.4	36.6
Cluster 8	10.0	90.0
Cluster 9	33.2	66.8
Cluster 10	31.7	68.3

2.4.3.1 Relevant clusters enriched in the CRPs that analyze very connected proteins

Interestingly, groups with high *dex* and *local_nE* values [clusters 1 and 2 with 2.3e-05 and 5.5e-07 binomial p-values respectively in SOM (Figure 4.5) and clusters 4, 5, 6, and 8 with 1.5e-03, 1.5e-19, 1.1e-04, and 1.9e-13 binomial p-value respectively in K-Means (Figure 4.6)] are enriched in CRPs (see Figure 4.5b and 4.6b). Tables 4.2 and 4.3 contain the percentages of CRP and NRPs within each cluster with SOM and K-Means strategies accordingly.

To further validate this observation associated with CRPs, we have studied how proteins are classified considering cancer cases and normal controls. In consequence, protein distribution in cancer and normal control cases are also analyzed on the space of *dex* and *local_nE* features (Figure 4.7). In the same direction, groups with the highest *local_nE* and *dex* values are significantly enriched on CRPs (clusters 1, 3, 7, and 10 with 5.3e-23, 1.1e-12, 5.9e-23, and 3.5e-14 binomial p-values respectively). Differently, clusters 6 and 9 are enriched on normal related proteins with 7.8e-29 and 7.8e-31 binomial p-value mutually.

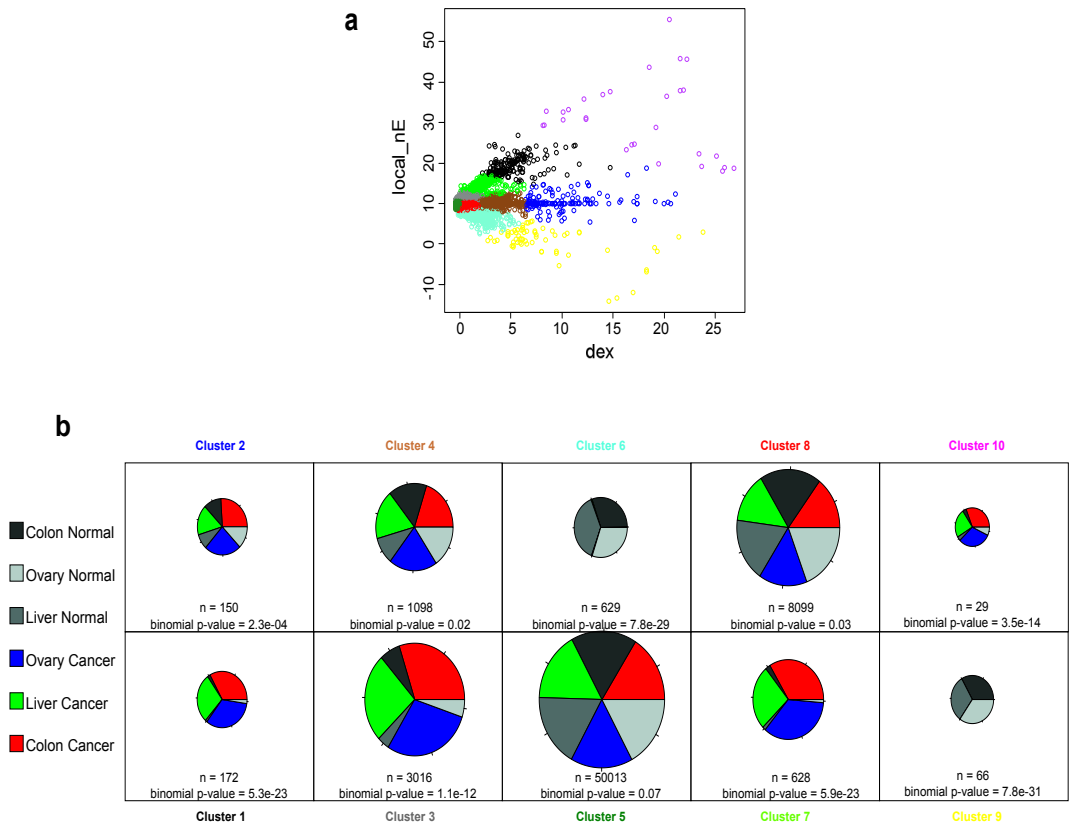


Figure 4.7: K-Means clustering analysis in cancer and normal control proteins (a) Representation of the proteins by their corresponding *dex* and *local_nE* scores after K-Means clusterization. (b) Pie charts containing the colored fraction of diseases in each cluster characterized by proteins with similar *local_nE* and *dex* values. Each fraction in the pie chart symbolizes the number of proteins in each disease with the feature values of the corresponding cluster. The *n* parameter represents the number of proteins in each cluster, and the p-value is the associated score with a binomial distribution.

2.4.3.2 Relevant clusters enriched in the NRPs that analyze less connected proteins

Conversely, groups with lower *dex* and *local_nE* values [clusters 7, 9 and 10 with 0.06, 2.3e-04 and 9.8e-06 binomial p-values respectively in SOM (Figure 4.5), and cluster 7 with 0.02 binomial p-value in K-Means (Figure 4.6)] are enriched in the NRPs (Table 4.2 and 4.3 specifically). In addition, these clusters correspond to the largest groups, containing almost half of the total amount of proteins included in the study. In particular, K-Means cluster 7 (Figure 4.6) is the one with the lowest *dex* and *local_nE* values, and it is significantly enriched in the NRPs, with the 63,4% of the proteins associated with a CNS disorder. Furthermore, cluster 7 is the largest cluster with 32,235 proteins, which corresponds with more than a half of the total amount of proteins included. Different from the CRPs, few NRPs are included in groups with high *dex* and *local_nE* values, being gathered the majority in low feature values groups.

To validate the less connectivity pattern in NRPs, we have studied how proteins are classified considering neurological cases and their corresponding normal controls. Similarly, NRPs are assembled in clusters with the lowest *dex* and *local_nE* values (clusters 1, 3 and 10 with 5.6e-15, 7.8e-31 and 3.1e-24 binomial p-values respectively). Oppositely clusters 4, 7 and 8 are enriched significantly (7.8e-29, 9.4e-22 and 3.9e-27 mutually) in normal related proteins, representing higher feature values (Figure 4.8).

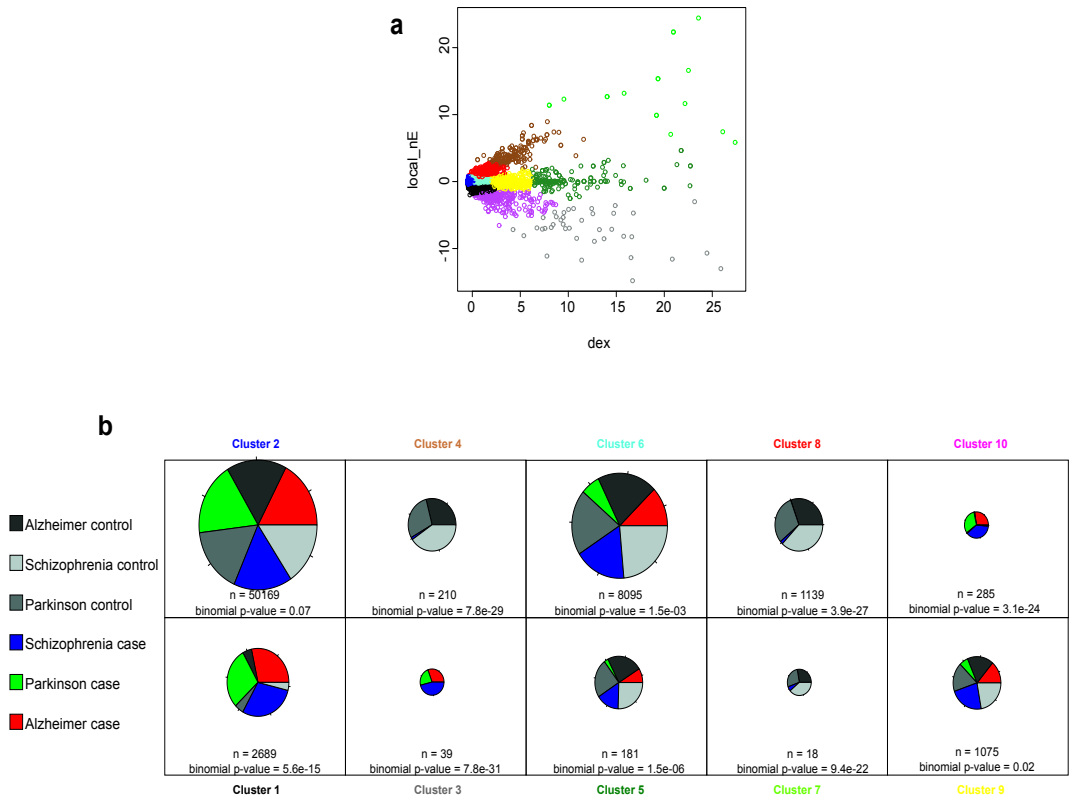


Figure 4.8: K-Means clustering analysis in CNS disorder and normal control proteins (a) Proteins by their corresponding *dex* and *local_nE* scores after K-Means clusterization. (b) Pie charts containing the colored fraction of diseases in each cluster characterized by proteins with similar *local_nE* and *dex* values.

4.5.4 Biological outcome

Following the proposed pipeline, once the clusters are obtained, we have analyzed the biological context behind the classification. To achieve this objective, an enrichment analysis on the biological pathways has been conducted using GeneCodis (gen; Carmona-Saez et al., 2007; Nogales-Cadenas et al., 2009; Tabas-Madrid et al., 2012) for each cluster of proteins. We have conducted in GeneCodis an enrichment analysis on KEGG pathways of the very connected proteins, and we have observed that these proteins are significantly enriched in pathways in cancer, focal adhesion, MAPK signaling pathway, Chemokine signaling pathway, and diverse inflammation pathways among others. In particular, the deregulation of the p53 signaling pathway is associated with the initiation and progression of cancer. Interestingly, recent studies previously point to a role for this pathway in CNS disorders (Tabarés-Seisdedos and Rubenstein, 2013).

The *PIN1* gene has been proposed to be a putative link between the pathogeneses of cancer and AD (Behrens et al., 2009). Figure 4.9 shows the different scenarios in Pin1 protein in ovarian cancer and AD. More than 30% of all the Pin1 interacting partners are expressed and active (red curves) when the ovarian samples are studied. However, *PIN1* gene in AD is not expressed, and hence, no active connections are found in AD (blue curves). This protein is associated with cell division (Lu, 2004) and is typically over-expressed in different cancers. Moreover, *PIN1* is depleted in AD. Different studies support the theory that it restores the function of the phosphorylated tau protein (Lu, 2004), and mouse models in which PIN1 is knocked down present neurodegenerative phenotypes (Liou et al., 2011).

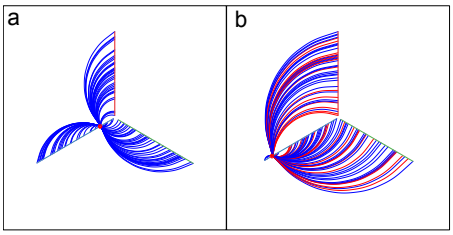


Figure 4.9: Neighbor interactions in Pin1 in (a) Alzheimer's disease and (b) ovarian cancer. The red Bezier curves represent the active interactions (Equation 1, case 2) and the blue Bezier curves represent the inactive interactions (Equation 1, cases 2 and 3). The hub corresponds to the Pin1 protein.

4.6 Discussion

A pattern recognition system is present, in which proteins related with CNS disorders (NRPs) and cancer related proteins (CRPs) are classified. We have defined two attributes (*dex* and *local_nE*) that are capable to categorize in different groups the majority of the proteins related to cancer or CNS disorders, corresponding with clusters with high or low feature values. The *dex* feature is indicative of the number of neighbors of each protein being expressed, important to estimate how connected and active proteins are within the PPIN. And we have adopted the *local_nE* attribute from our previous work (Ibáñez et al., 2015), which measures the stability of a single protein in its neighborhood.

We have validated that the definition of these two features are adequate to classify the proteins into CRP and NRPs, according to them. Successful classifications have been obtained with Random Tree Forest, J48 and Bayes classifiers. We have performed two different classification approaches, SOM and K-Means, to validate that our results do not depend on the methodology, reinforcing the result.

The first significant finding in this study is that protein groups characterized by high *dex* and *local_nE* are significantly enriched in CRPs. This means, that proteins within these kinds of clusters are well connected and active within the network, and so are their neighbors. In these clusters, the majority of proteins correspond to codifying genes highly expressed in ovarian, colon or liver cancer. This observation goes in the same direction as previous work, in which studies of the network topology shows that cancer proteins tend to be central within the network, being highly connected (Jonsson and Bates, 2006; Taylor et al., 2009; Sun and Zhao, 2010; Xia et al., 2011; Choura and Rebaï, 2012; Xiong et al., 2014). Others such as Jeong et al. (2001), propose that important proteins for a cell's survival are highly connected and altering them might have profound effects on the interaction network. According to the definition of the *local_nE*, higher values in this feature represent active connections between nodes of expressed genes, indicating that any alteration in this node will destabilize the network.

Moreover, we have validated further this consideration analyzing cancer cases and normal controls proteins with similar outcome.

Our second important finding is that conversely, protein groups characterized by low *dex* and *local_nE* are significantly enriched on NRPs. Proteins inside these clusters are less connected and active within the network, an so their neighbors. Furthermore, when proteins in CNS disorder cases and normal controls are studied, proteins which codifying genes are highly expressed in normal cases, correspond majority to the groups with high scores of *dex* and *local_nE*.

Our approach based on the properties of the network (*dex* and *local_nE*) appears to be capable of identifying candidate protein groups potentially associated with cancer or CNS disorders. The clusters of proteins with higher *dex* and *local_nE* values tend to be composed of cancer-related genes, and those with moderate *dex* and *local_nE* values are usually composed of neurological-related genes. In fact, *dex* and *local_nE* attributes may be considered as a hallmark of the topological properties of the CRP and NRPs.

These findings are interesting from the point of view of comorbidity studies. Tabarés-Seisdedos and Rubenstein (2013) and Tabarés-Seisdedos et al. (2011), among others, put forward the hypothesis that some CNS diseases could affect the risk of developing some types of cancer. Many reviews have been published on this topic (Tabarés-Seisdedos and Rubenstein, 2013). Meta-analyses on direct and inverse cancer comorbidity in people with CNS diseases have also been conducted.

5 Conclusions and future work

5.1 General conclusions

Data mining and machine learning techniques are the drivers of this work. We have proved here that large amount of biological and medical data can be analyzed integrating advanced computational methods. In the context of the biomedical theme of *inverse comorbidity*, we have presented in chapters 2, 3, and 4 three computational approaches with different purposes. The results of the application of these proposed strategies show how the application of new methodologies based on data mining and machine learning techniques can produce progress in specific domains of biomedicine.

We have analyzed global genome wide transcriptomic data together with a meta-analysis strategy, and we have further studied complex relations in biological systems beyond the functioning of the components in isolation. We have investigated the use of information on interaction networks (i.e., protein interactions) in order to follow better the relations in complex systems behavior, and to study dynamic aspects. The use and application of classification methods to assemble specific proteins of cancer or CNS disorders has turned out to be a relevant approach in the context of this thesis.

In terms of Biology, we have identified for the first time a set of genes and pathways deregulated in opposite directions in three CNS disorders and three cancer types. This finding sets the molecular basis of previously described *inverse comorbidity* pattern observed at the population level. The identification of those specific genes and pathways will provide the first clue for follow-up experimental approaches, including their investigation of their potential relevance as therapeutic strategies.

To our knowledge, this work represents the first systematic attempt to the identification of possible molecular base of *inverse comorbidity* associations.

In a follow-up study analyzing the addition of information on protein–protein interaction networks, allowed us to show that cancer related networks are more unstable than networks associated with neurological disorders. Finally, in a third study, using SOM, K-Means, and classification systems we have defined two features that have allowed us to categorize and sort sets of proteins characteristics of each disorder.

This thesis represents a collaboration between bioinformatics, computational biology, and artificial intelligence, in which computational methods based on machine learning techniques have been proposed to better understand the biological context behind the *inverse comorbidity*. In previous chapters methodologies, results, and conclusions corresponding to these three studies have been presented. In this chapter the principal conclusions are summarized, and future perspectives described.

5.2 Development of a data mining approach to perform transcriptomic meta-analyses between cancer–CNS disorders

A relevant number of epidemiological studies have demonstrated a lower-than-expected probability of developing some types of cancer in patients with certain CNS disorders (known as *inverse comorbidity*). This behavior suggests that *inverse comorbidity* may be influenced by environmental factors, drugs treatments and other aspects related with disease diagnosis. Genetics is at the basis of this behavior, and it regulates and influences to this contribution in combination with external factors. This adaptation represents an open door to understand why certain individuals are protected against many different types of cancer. For such aim, our principal objective was to discover the molecular mechanisms that underlie this apparent protective effect.

In the first part in this thesis and described in chapter 2, we present a novel data mining approach in order to compute transcriptomic meta-analyses between some types of cancer and CNS disorders. We hypothesize that a molecular substrate exists and is shared in a specific manner in cancer and CNS disorders. Accordingly, we propose here the deregulation in opposite directions of a common set of genes and pathways as an underlying cause of *inverse comorbidity*. Using gene expression data, our principal idea was to test whether there is a significant overlap between genes that are up-regulated in cancer and down-regulated in CNS disorders, and the other way around.

The significant result here is that a significant overlap is observed between the genes up-regulated in cancers and down-regulated in CNS disorders, and conversely, between the genes down-regulated in cancers and up-regulated in CNS disorders. These gene expression deregulations in opposite directions are also observed at the level of pathways, and point to specific genes and functions the deregulation of which could promote CNS disorders and simultaneously lowers the initiation or progression of cancer.

In the future, further analyses will be necessary to conclude to a direct protective effect of gene expression deregulations in cancer-prone tissues of patients suffering from CNS disorders. This identification of antagonistically deregulated genes and pathways in complex diseases that have been previously described as inversely comorbid provides, to our knowledge, the first systematic insights into the possible molecular base of these associations, acknowledging previous important works (Behrens et al., 2009). It suggests that the up-regulation of a set of genes or processes could increase the incidence of CNS disorders and simultaneously lower the chances of developing cancers, while the down-regulation of another set of genes or processes could contribute to a decrease in the incidence of CNS disorders while increasing the cancer risks.

The post-mortem brain samples in CNS disorders have likely received drug treatments. Hence, the observed expression deregulations could be the consequence of the drugs administered to the patients. If this is the case, it can be hypothesized that some of the drugs used to treat CNS disorders might be able to revert the expression of a number of cancer genes. In this context, the repurposing of drugs from the CNS to the cancer field could open new therapeutic avenues. Indeed some punctual observations

have been made. For example, a drug repositioning bioinformatic approach have identified that triclyclic antidepressant and related molecules could potently induce apoptosis in small cell lung cancer and other neuroendocrine tumors (Jahchan et al., 2013). Another example is the thioridazine, an anti-psychotic drug antagonizing the dopamine receptor and potentially able to alter physiological states and expression patterns, have been reported to target cancer stem cells selectively (Sachlos et al., 2012). In the other direction, there is another work presented in the section 2.6, in which they have proved that a failed drug on treating solid tumor appears to restore synaptic connections and reduced inflammation, and the animal's memory, a hallmark of Alzheimer's disease. In this manner, memory and the connections between brain cells are restored in mice with a model of Alzheimer's given this experimental cancer drug (Kaufman et al., 2015).

Despite these two last observations, the effect of the drugs cannot explain by themselves the observed *inverse comorbidity*. For instance, several works have noted that the relatives of patients suffering schizophrenia have less probability of developing any cancer (Gal et al., 2012; Ji et al., 2013), suggesting that genes associated with schizophrenia might confer reduced cancer susceptibility. There is a genomic and molecular base that determines the behavior of the individual in general, and the effect of the drug in each one, in particular. More data is necessary to be able to establish a relationship between some drugs and the genetics behind the *inverse comorbidity*, and the observed gene expression patterns.

5.3 Development of a machine learning approach inspired by simulated annealing to study the stability of PPINs in cancer-CNS disorders

Following the molecular basis underlying the differences between cancers and neurological conditions, we have integrated here gene expression data with protein–protein interaction networks to study these differences in terms of network organization rather than at the level of individual genes. Molecular networks, and in particular, protein–protein interaction networks provide a powerful tool for the study of biomedical systems. Diseases can alter the structure of the network (Ideker and Sharan, 2008), and it has also be pointed out that they can also alter the stability of networks (Teschendorff and Severini, 2010; West et al., 2012), and studying network stability might be

fundamental to the better understanding of the biological systems behave, apart from components in isolation. Stability describes a network state that is not significantly altered, even when fundamental properties have changed or perturbations have been introduced.

In the second part in this thesis and described in chapter 3, we have designed an approach inspired on SA, representing PPINs as systems of nodes that are dynamically updated towards a global state of stability. Our strategy is based on the definition of a neighbor-energy function (*nE*) that measures the stability of the network in the general deterministic approach, where *nE* indicates network stability, and it can be interpreted in terms of resistance to alterations and perturbations. We have analyzed a large set of experimental data on gene expression and various PPINs.

The first significant finding of this study is that networks containing information about expression in four human cancers (ovarian, colon, kidney, and liver) are less stable than the control networks of normal samples. Moreover, this instability in the network seems to increase as these cancers evolve, at least in the tumor progression data sets analyzed. The approach employed is based on the analyses of samples in different conditions and it does not include temporal evolution per se. Thus, the results obtained by analyzing the temporal progression of tumors can be taken as an indication of network evolution towards a less stable state, and a way of reconciling our methodology with the standard SA applications.

The second important finding is that the AD network is more stable than the corresponding control normal network, with a significant increase in the *nE* of the corresponding networks. This is an interesting behavior that contrasts with that of cancers, and as far as we know is detected here for the first time. One possible interpretation of these results would be that cancer implies a general deregulation of cell growth through the hyper-activation of certain pathways, resulting in a destabilization of their interactions, while AD and other neurological disorders imply the stabilization of biological processes and network interactions, and their general slowing down. The striking contrast in the behavior of cancer and AD networks, from less to more stable networks, should be considered in the context of the observed *inverse comorbidity* of these two groups of diseases.

5.4 Development of a pattern recognition method for the recognition between cancer–CNS disorders related proteins

Studies of the topology of PPINs show that cancer proteins tend to be central within the network as they are highly connected (Jonsson and Bates, 2006; Taylor et al., 2009; Sun and Zhao, 2010; Xia et al., 2011; Choura and Rebaï, 2012; Xiong et al., 2014). It has also been proposed that important proteins for a cell's survival are highly connected (Jeong et al., 2001), and altering them has profound effects on the interaction network. Similarly, studies on the genomic and network characteristics of genes that mutated in cancer indicate that these genes tend to encode central hubs within a PPIN (Rambaldi et al., 2008).

In the third section in this thesis and described in chapter 4, following the main hypothesis and results shown in chapters 2 and 3, we examine critically whether cancer related proteins (CRPs) tend to be more connected. Indeed, we go further, and we also analyze whether proteins related with neurological disorders (NRPs) have the opposite behavior than CRPs, pursuing the *inverse comorbidity* theory in these complex disorders. A pattern recognition system is designed for the classification of NRPs and CRPs. For such objective, we have defined two attributes (*dex* and *partial_nE*) that are capable to categorize in different groups the majority of the proteins related to cancer or CNS disorders, corresponding with clusters with high or low feature values. The *dex* feature is indicative of the number of neighbors of each protein being expressed, important to estimate how connected and active proteins are within the PPIN. And we have adopted the *local_nE* attribute from the previous study [chapter 3, (Ibáñez et al., 2015)], which measures the stability of a single protein in its neighborhood.

We have validated that the definition of these two features are adequate to classify the proteins into CRP and NRPs, according to them. Successful classifications have been obtained with Random Tree Forest, J48, and Bayes classifiers. We have performed two different classification approaches, SOM and K-Means, to validate that our results do not depend on the methodology, reinforcing the result.

The first significant finding in this study is that protein groups characterized by high *dex* and *partial_nE* are significantly enriched in CRPs. This means, that proteins within these kinds of clusters are well connected and active within the network, and so are their neighbors. In these clusters, the majority of proteins correspond to codifying genes highly expressed in ovarian, colon or liver cancer. This observation goes in the same direction as previous work, in which studies of the network topology shows that cancer proteins tend to be central within the network, being highly connected (Jonsson and Bates, 2006; Taylor et al., 2009; Sun and Zhao, 2010; Xia et al., 2011; Choura and Rebaï, 2012; Xiong et al., 2014). Others such as Jeong et al. (2001), propose that important proteins for a cell's survival are highly connected and altering them might have profound effects on the interaction network. According to the definition of the *local_nE*, higher values in this feature represent active connections between nodes of expressed genes, indicating that any alteration in this node will destabilize the network. Moreover, we have validated further this consideration analyzing cancer cases and normal controls proteins with similar outcome.

Our second important finding is that conversely, protein groups characterized by low *dex* and *local_nE* are significantly enriched on NRPs. Proteins inside these clusters are less connected and active within the network, an so their neighbors. Furthermore, when proteins in CNS disorder cases and normal controls are studied, genes highly expressed (which codify proteins) in normal cases, are majority in groups with high scores of *dex* and *partial_nE*.

Our system based on the properties of the network (*dex* and *local_nE*) appears to be capable of identifying candidate protein groups potentially associated with cancer or CNS disorders. The clusters of proteins with higher *dex* and *local_nE* values tend to be composed of cancer-related genes, and those with moderate *dex* and *local_nE* values are usually composed of neurological-related genes. In fact, *dex* and *local_nE* attributes may be considered as a hallmark of the topological properties of the CRP and NRPs. These findings are interesting from the point of view of comorbidity studies.

5.5 Future directions

The analysis of opposite expression deregulations in cancer and CNS disorders can be considered as an initial step toward a more exhaustive understanding of the *inverse comorbidity*, and the approach we here present could serve as a new strategy to investigate further possible relations between other complex diseases [for a disease comorbidities review see (Catalá-López et al., 2014b)]. Further analyses will be necessary to see whether this kind of gene expression relations take part in more future available data. Regarding the protective effect, since gene regulation change the phenotype, it could explain the protection against a disease; but additional complex and systematic experiments should be done.

The computational approach inspired on the DSA to analyze PPIN stability in complex disorders could be used in the characterization of many samples in cancer evolution, and validate it further with a clinical confirmation step (i.e., correlate the cancer stage with the corresponding sample *nE*, according to the nE evolution within that cancer study). It could also be used as a classifier to distinguish cancer and normal samples. Another possibility will be to cluster the results of this procedure in order to extract specific proteins for which additional experimental information could be available. Furthermore, this scheme could be also applied to any network system where the elements are characterized by a state S_i and their interactions associated to a weight W_{ij}. In a biological context there are numerous systems with these characteristics, such as protein interaction and gene control networks. In particular, it could be interesting to apply into a protein–protein interaction network, in which different drugs are introduced, and analyze the system behaviors (defining properly the parameters).

The expert system for classifying proteins related with cancer and neurological disorders, could be used to classify and characterize proteins that are strongly conserved in cancer evolution from the point of view of the connectivity within the network. Moreover, it could be considered as an initial example toward the characterization of proteins related with other complex disorders, estimating how connected and active proteins are within the PPIN, as well as measuring the stability of each protein in its neighborhood. This strategy could help ranking a set of candidate proteins or genes, that might play a key role within certain complex disorders.

A Supplementary Figures

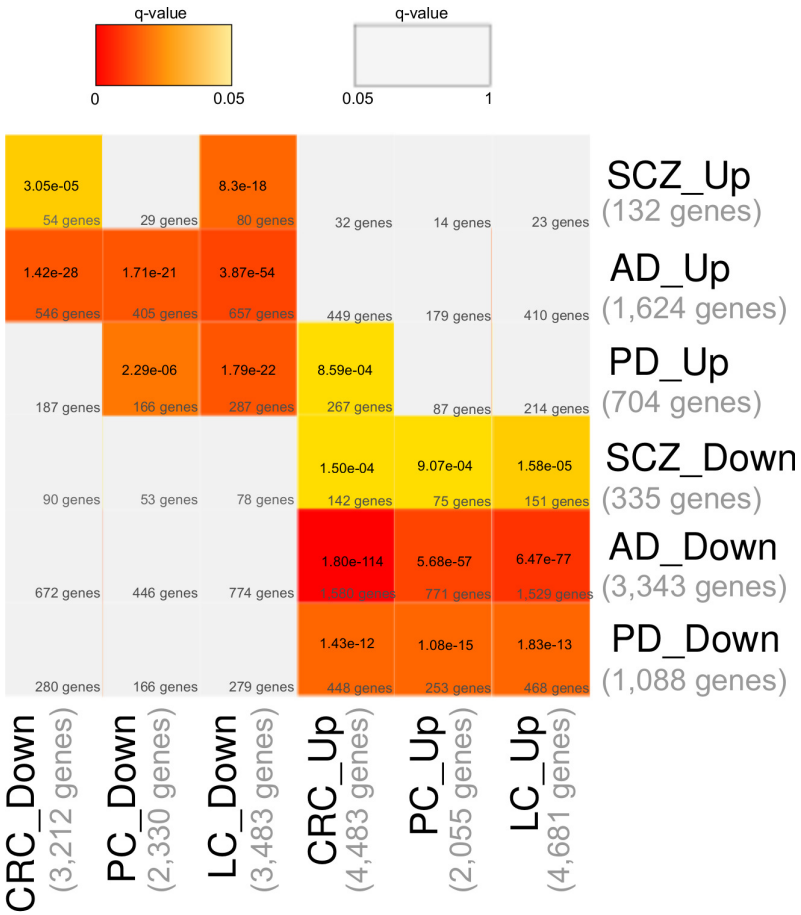


Figure A.1: Comparisons of DEGs associated with CNS disorders and cancers at 0.005. The DEGs up- and down-regulated after gene expression meta-analysis in each CNS disorder (Alzheimer's Disease, AD; Parkinson's Disease, PD; and Schizophrenia, SCZ) and in each Cancer (Colorectal Cancer, CRC; Prostate Cancer, PC; Lung Cancer, LC) are compared to each others.

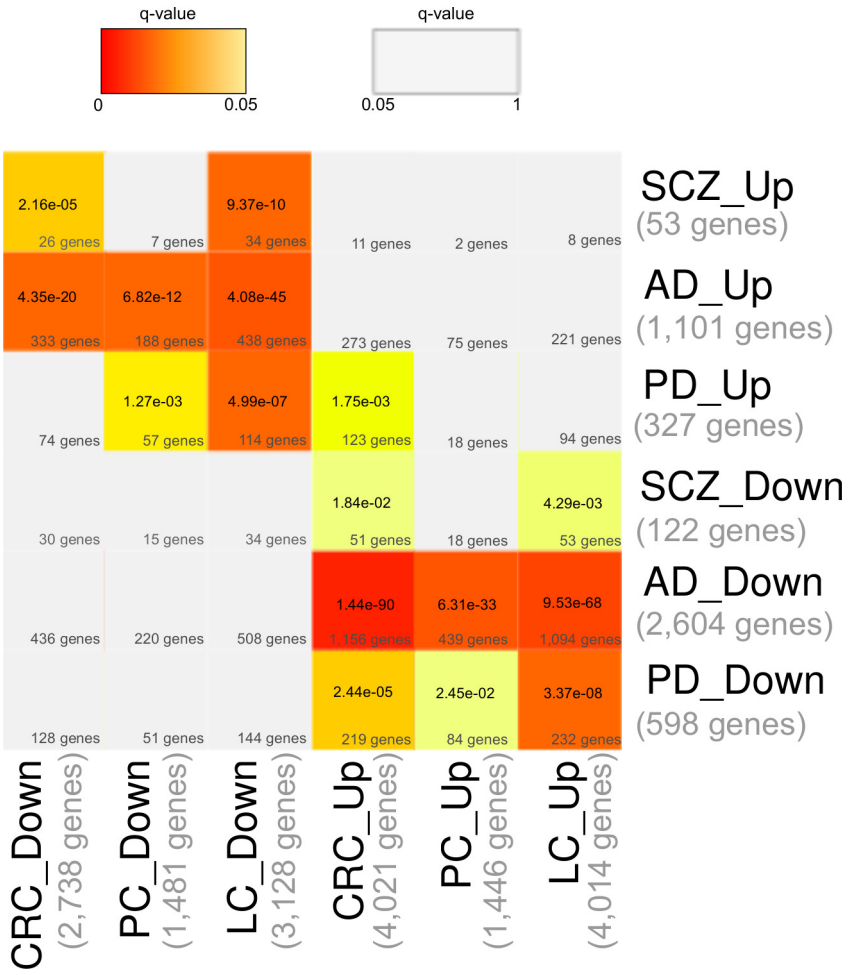


Figure A.2: Comparisons of DEGs associated with CNS disorders and cancers at 0.0005. The DEGs up- and down-regulated after gene expression meta-analysis in each CNS disorder (Alzheimer's Disease, AD; Parkinson's Disease, PD; and Schizophrenia, SCZ) and in each Cancer (Colorectal Cancer, CRC; Prostate Cancer, PC; Lung Cancer, LC) are compared to each others.

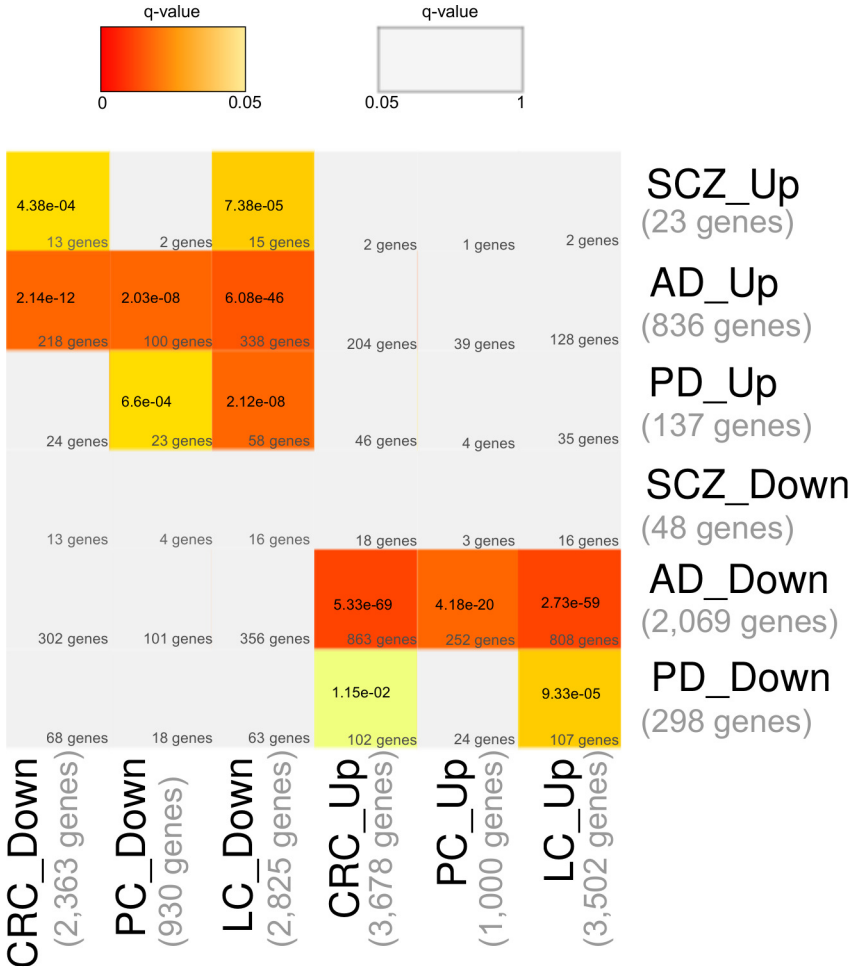


Figure A.3: Comparisons of DEGs associated with CNS disorders and cancers at 0.00005. The DEGs up- and down-regulated after gene expression meta-analysis in each CNS disorder (Alzheimer's Disease, AD; Parkinson's Disease, PD; and Schizophrenia, SCZ) and in each Cancer (Colorectal Cancer, CRC; Prostate Cancer, PC; Lung Cancer, LC) are compared to each others.

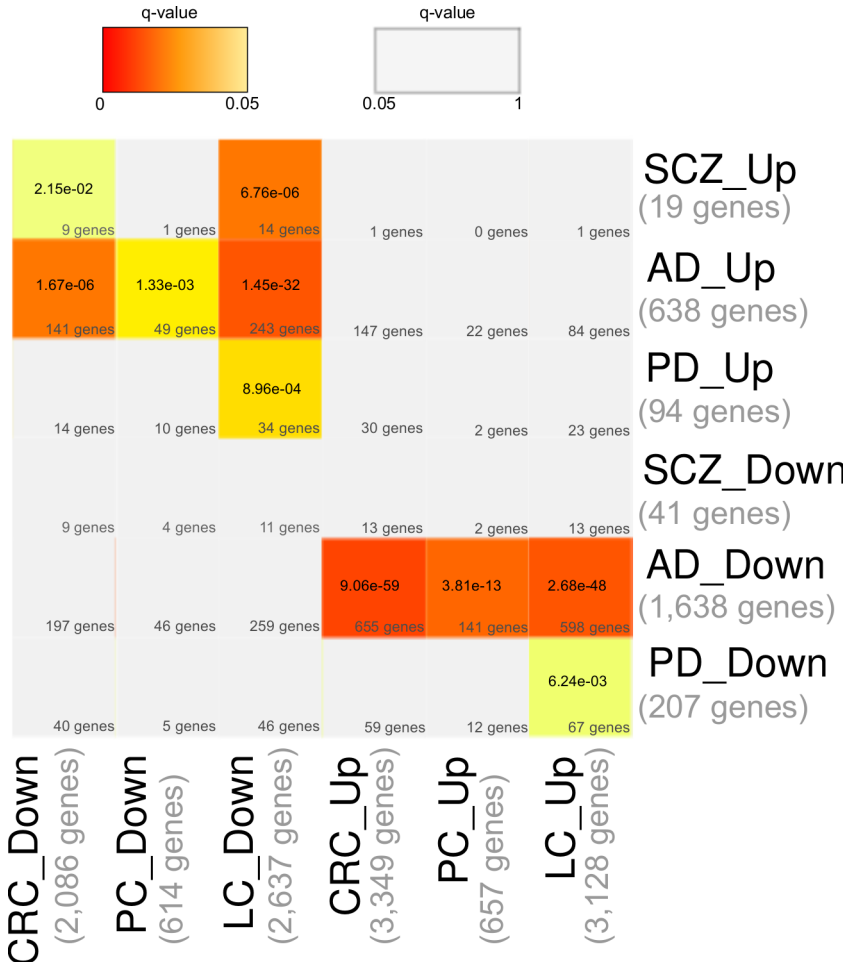


Figure A.4: Comparisons of DEGs associated with CNS disorders and cancers at 0.000005. The DEGs up- and down-regulated after gene expression meta-analysis in each CNS disorder (Alzheimer's Disease, AD; Parkinson's Disease, PD; and Schizophrenia, SCZ) and in each Cancer (Colorectal Cancer, CRC; Prostate Cancer, PC; Lung Cancer, LC) are compared to each others.

Appendix A. Supplementary Figures

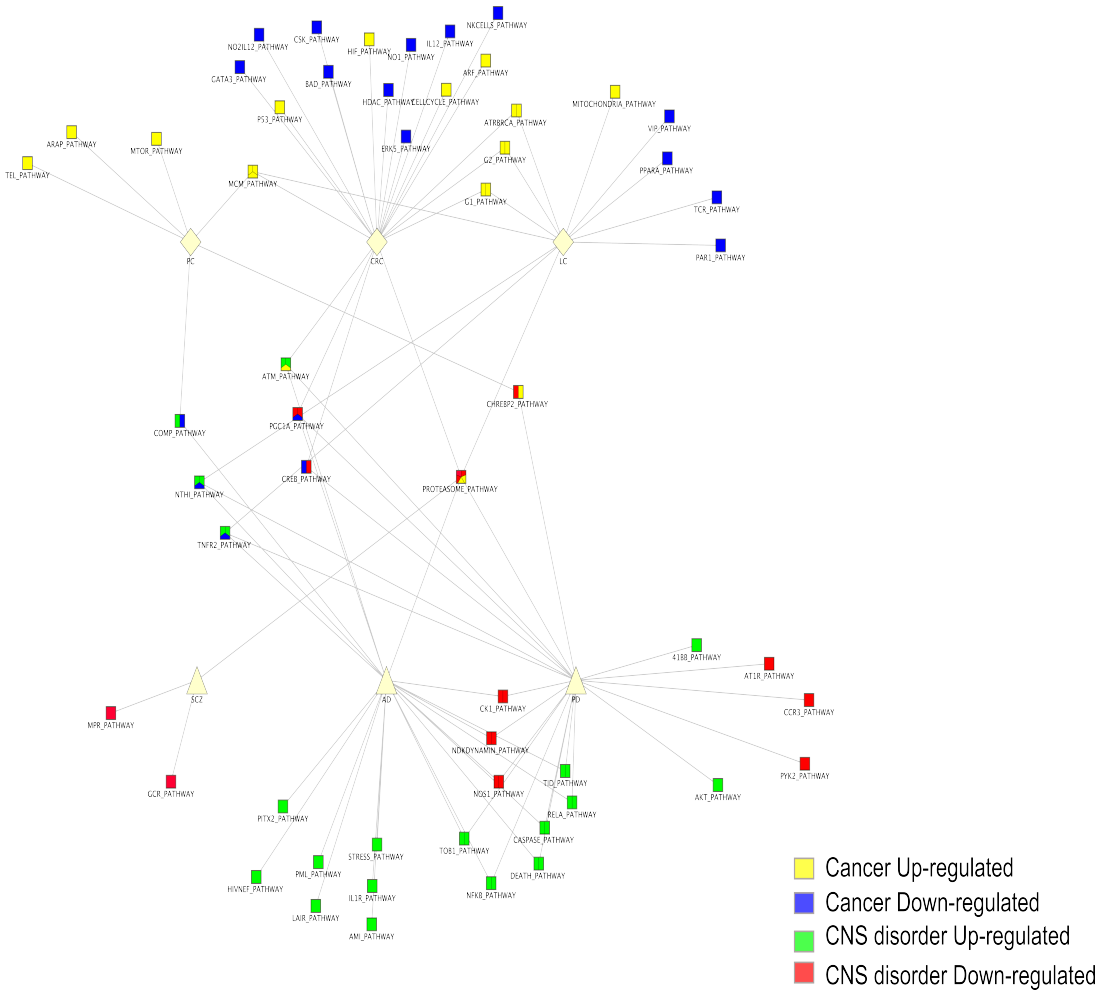


Figure A.5: Biocarta pathway significantly deregulated in the three types of cancers and CNS disorders. Cancer upregulated (yellow), cancer downregulated (blue), CNS disorder upregulated (green) and CNS disorder downregulated (red). The green/blue and yellow/red associations thus correspond to pathways deregulated in opposite directions in CNS disorders and cancers.

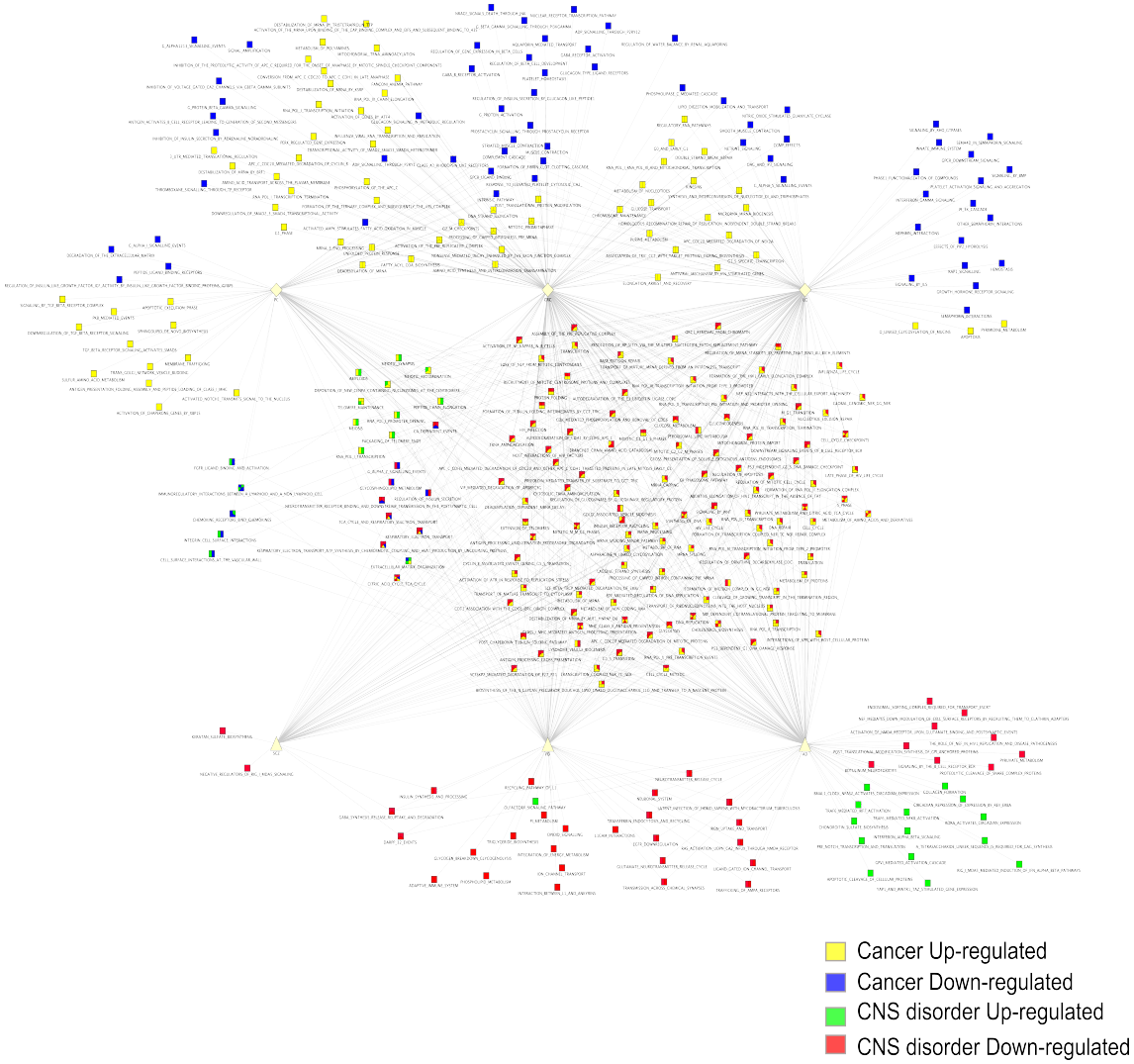


Figure A.6: Reactome pathway significantly deregulated in the three types of cancers and CNS disorders. Cancer upregulated (yellow), cancer downregulated (blue), CNS disorder upregulated (green) and CNS disorder downregulated (red). The green/blue and yellow/red associations thus correspond to pathways deregulated in opposite directions in CNS disorders and cancers.

B Supplementary material - Network stability study

B.1 Description of the deterministic simulated annealing algorithm

Begin $initialize\ nE,\ W_{ij},\ S_i,\ S_j,\ i,\ j = 1..N\ t = 0$

do $t = t + 1$

$nE = -\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} * Si * S_j$

$t = t + 1$ **until** $t = tmax$ **end**

whereas:

nE is the system's final energy. t corresponds to a gene expression sample in each dataset and type (Normal, Cancer, Neurological). W_{ij} describes the weight explained in Equation 3.1, representing the existing influence between nodes S_i and S_j. In our approach S_i is the significance level of the expression or no expression of each gene.

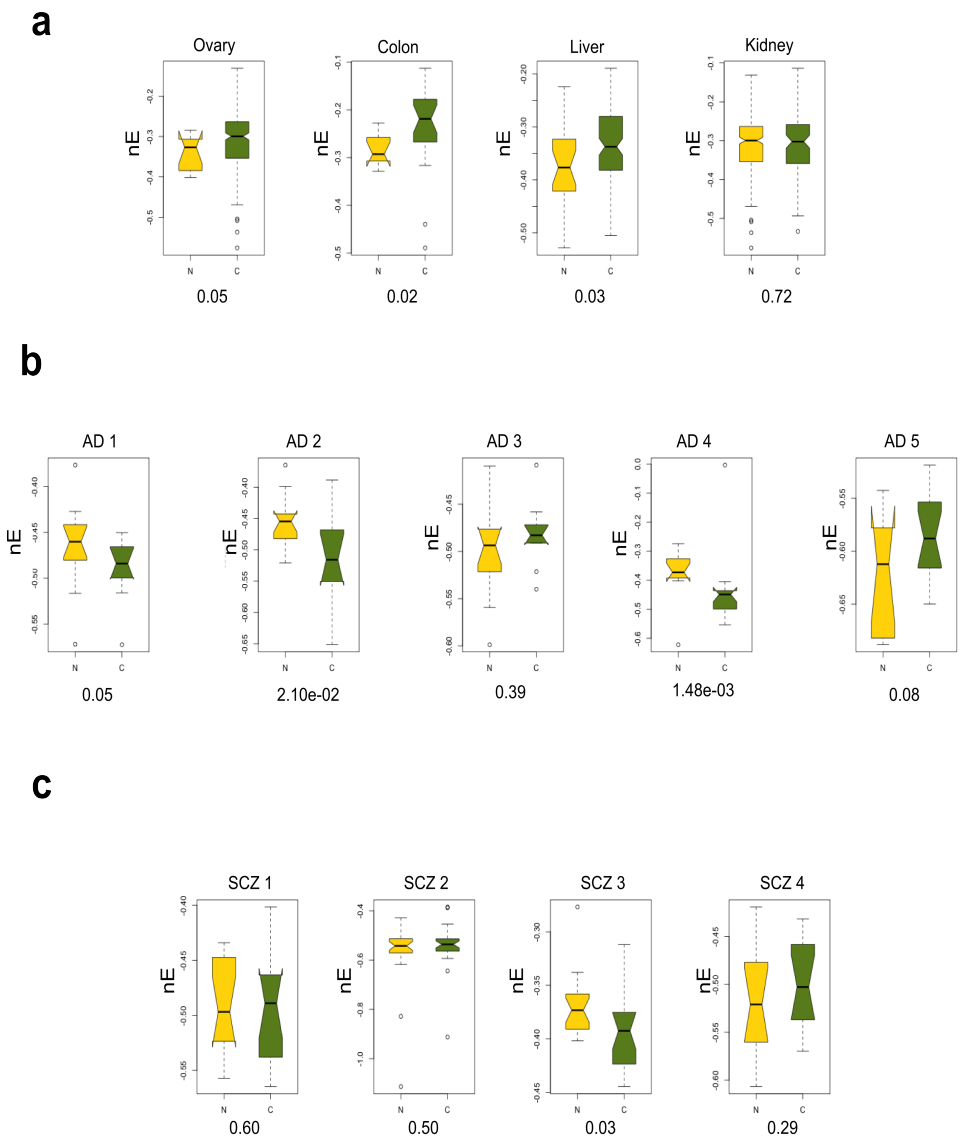


Figure B.1: The *nE* distribution that maps all the genes in the HPRD network in the: (a) normal (N) and cancer (C) states (Ovarian, Colon, Liver and Kidney); (b) Normal (N) and AD (C); (c) Normal (N) and SCZ disease (C) state. The Wilcoxon-rank p-value is presented below the x-axis.

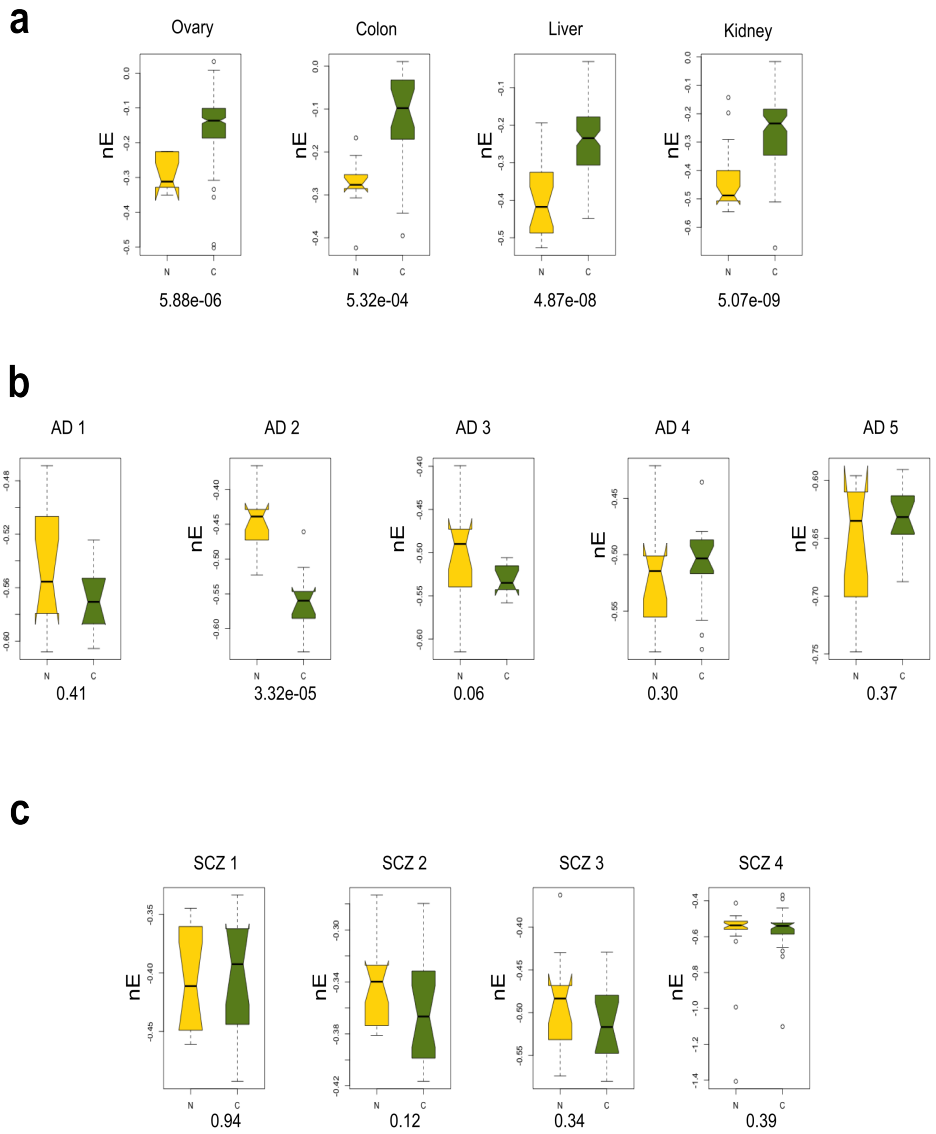


Figure B.2: The *nE* distribution that maps all the genes in the HIPPIE network in the: (a) normal (N) and cancer (C) states (Ovarian, Colon, Liver and Kidney); (b) Normal (N) and AD (C); (c) Normal (N) and SCZ disease (C) state. The Wilcoxon-rank p-value is presented below the x-axis.

Bibliography

Biocarta. http://www.biocarta.com/genes/index.asp. Online, accessed February 2013.

Ae - array express. http://www.ebi.ac.uk/arrayexpress/. Online, accessed 2011-07-21.

Gene expression barcode. http://barcode.luhs.org/. Online, accessed 2011-10-21.

clustersim r package. http://CRAN.R-project.org/package=clusterSim. Online.

fpc r package. http://CRAN.R-project.org/package=fpc. Online.

Genecodis. http://genecodis.cnb.csic.es. Online, accessed on March 2015.

Ncbi geo - gene expression omnibus. http://www.ncbi.nlm.nih.gov/geo. Online, accessed 2011-07-21 and 2011-10-21.

Hippie (protein-protein interaction network). http://cbdm.mdc-berlin.de/tools/hippie/download.php. Online, accessed 2014-09-05.

Hprd (protein-protein interaction network). http://www.hprd.org/download. Online, accessed 2014-09-03.

Kegg pathway - synaptic vesicle cycle. http://www.genome.jp/dbget-bin/www_bget?pathway:hsa04721. Online, accessed 2014-09-25.

Pina (protein-protein interaction network). http://cbg.garvan.unsw.edu.au/pina/interactome.stat.do. Online, accessed 2011-10-19.

Ncbi pubmed. http://www.ncbi.nlm.nih.gov/pubmed. Online, accessed 2014-11-16.

Smri - stanley medican research institute. https://www.stanleygenomics.org. Online, accessed 2012-03-15.

som r package. http://CRAN.R-project.org/package=som. Online.

stats r package. https://stat.ethz.ch/R-manual/R-patched/library/stats/html/kmeans.html. Online.

The cancer gene atlas. https://tcga-data.nci.nih.gov/tcga/. Online.

Current topics in computational molecular biology. In *MIT Press.*, 2002.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000. ISSN 0028-0836. doi: 10.1038/35000501.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.*, 96(12):6745–50, 1999. ISSN 0027-8424.

Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997. ISSN 13624962. doi: 10.1093/nar/25.17.3389.

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., Kellis, M., MacArthur, D. G., Getz, G., Shabalin, A. A., Li, G., Zhou, Y.-H., Nobel, A. B., Rusyn, I., Wright, F. A., Lappalainen, T., Ferreira, P. G., Ongen, H., Rivas, M. A., Battle, A., Mostafavi, S., Monlong, J., Sammeth, M., Mele, M., Reverter, F., Goldmann, J. M., Koller, D., Guigo, R., McCarthy, M. I., Dermitzakis, E. T., Gamazon, E. R., Im, H. K., Konkashbaev, A., Nicolae, D. L., Cox, N. J., Flutre, T., Wen, X., Stephens, M., Pritchard, J. K., Tu, Z., Zhang, B., Huang, T., Long, Q., Lin, L., Yang, J., Zhu, J., Liu, J., Brown, A., Mestichelli, B., Tidwell, D., Lo, E., Salvatore, M., Shad, S., Thomas, J. A., Lonsdale, J. T., Moser, M. T., Gillard, B. M., Karasik, E., Ramsey, K., Choi, C., Foster, B. A., Syron, J., Fleming, J., Magazine, H., Hasz, R., Walters, G. D., Bridge, J. P., Miklos, M.,

Sullivan, S., Barker, L. K., Traino, H. M., Mosavel, M., Siminoff, L. A., Valley, D. R., Rohrer, D. C., Jewell, S. D., Branton, P. A., Sobin, L. H., Barcus, M., Qi, L., McLean, J., Hariharan, P., Um, K. S., Wu, S., Tabor, D., Shive, C., Smith, A. M., Buia, S. A., Undale, A. H., Robinson, K. L., Roche, N., Valentino, K. M., Britton, A., Burges, R., Bradbury, D., Hambright, K. W., Seleski, J., Korzeniewski, G. E., Erickson, K., Marcus, Y., Tejada, J., Taherian, M., Lu, C., Basile, M., Mash, D. C., Volpi, S., Struewing, J. P., Temple, G. F., Boyer, J., Colantuoni, D., Little, R., Koester, S., Carithers, L. J., Moore, H. M., Guan, P., Compton, C., Sawyer, S. J., Demchok, J. P., Vaught, J. B., Rabiner, C. A., and Lockhart, N. C. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).*, 348(6235):648–660, 2015. ISSN 0036-8075. doi: 10.1126/science.1262110.

Bajaj, A., Driver, J. A., and Schernhammer, E. S. Parkinson's disease and cancer risk: a systematic review and meta-analysis. *Cancer Causes Control*, 21(5):697–707, 2010. ISSN 1573-7225. doi: 10.1007/s10552-009-9497-6.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, 12(1):56–68, 2011. ISSN 1471-0064. doi: 10.1038/nrg2918.

Barna, M., Pusic, A., Zollo, O., Costa, M., Kondrashov, N., Rego, E., Rao, P. H., and Ruggero, D. Suppression of Myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature*, 456(7224):971–5, 2008. ISSN 1476-4687. doi: 10.1038/nature07449.

Behrens, M. I., Lendon, C., and Roe, C. M. A common biological mechanism in cancer and Alzheimer's disease? *Curr. Alzheimer Res.*, 6(3):196–204, 2009. ISSN 1875-5828.

Behrens, M. I., Silva, M., Salech, F., Ponce, D. P., Merino, D., Sinning, M., Xiong, C., Roe, C. M., and Quest, A. F. G. Inverse susceptibility to oxidative death of lymphocytes obtained from Alzheimer's patients and skin cancer survivors: increased apoptosis in Alzheimer's and reduced necrosis in cancer. *J. Gerontol. A. Biol. Sci. Med. Sci.*, 67 (10):1036–40, 2012. ISSN 1758-535X.

Bell, D. S. Heart Failure: A Serious and Common Comorbidity of Diabetes. *Clin. Diabetes*, 22(2):61–65, 2004. ISSN 0891-8929. doi: 10.2337/diaclin.22.2.61.

Ben-Dor, A., Shamir, R., and Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.*, 6(3-4):281–97, 2004. ISSN 1066-5277. doi: 10.1089/106652799318274.

Bendhaiba, L., Boelaert, J., Olteanu, M., and Villa-Vialaneix, N. *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs*, 2015. R package version 1.0.

Bennett, M. K. and Scheller, R. H. The molecular machinery for secretion is conserved from yeast to neurons. *Proc. Natl. Acad. Sci. U. S. A.*, 90(7):2559–63, 1993. ISSN 0027-8424.

Bernardes, J. S. and Pedreira, C. E. A review of protein function prediction under machine learning perspective. *Recent Pat. Biotechnol.*, 7(2):122–41, 2013. ISSN 2212-4012.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J., Bellgard, M., Wassenaar, T. M., and Ussery, D. W. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics*, 6 (3):165–85, 2006. ISSN 1438-793X. doi: 10.1007/s10142-006-0027-2.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods*, 1 (2):97–111, 2010. ISSN 17592879. doi: 10.1002/jrsm.12.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. *Introduction to Meta-Analysis*. 2011.

Börnigen, D., Pers, T. H., Thorrez, L., Huttenhower, C., Moreau, Y., and Brunak, S. r. Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic Acids Res.*, 41(18):e171, 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt661.

Bredel, M., Scholtens, D. M., Yadav, A. K., Alvarez, A. A., Renfrow, J. J., Chandler, J. P., Yu, I. L. Y., Carro, M. S., Dai, F., Tagge, M. J., Ferrarese, R., Bredel, C., Phillips, H. S., Lukac, P. J., Robe, P. A., Weyerbrock, A., Vogel, H., Dubner, S., Mobley, B., He, X., Scheck, A. C., Sikic, B. I., Aldape, K. D., Chakravarti, A., and Harsh, G. R. NFKBIA deletion in glioblastomas. *N. Engl. J. Med.*, 364(7):627–37, 2011. ISSN 1533-4406. doi: 10.1056/NEJMoa1006312.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Bruggeman, F. J. and Westerhoff, H. V. The nature of systems biology. *Trends Microbiol.*, 15(1):45–50, 2007. ISSN 0966-842X. doi: 10.1016/j.tim.2006.11.003.

Cai, C., Langfelder, P., Fuller, T. F., Oldham, M. C., Luo, R., van den Berg, L. H., Ophoff, R. A., and Horvath, S. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*, 11(1):589, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-589.

Calinski, T. and Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods*, 3(1):1–27, 1974. ISSN 0361-0926. doi: 10.1080/03610927408827101.

Callaway, D. S., Newman, M. E., Strogatz, S. H., and Watts, D. J. Network robustness and fragility: percolation on random graphs. *Phys. Rev. Lett.*, 85(25):5468–71, 2000. ISSN 0031-9007.

Campain, A. and Yang, Y. H. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, 11:408, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-408.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., and Pascual-Montano, A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, 8(1):R3, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-1-r3.

Cason, A. L., Ikeguchi, Y., Skinner, C., Wood, T. C., Holden, K. R., Lubs, H. A., Martinez, F., Simensen, R. J., Stevenson, R. E., Pegg, A. E., and Schwartz, C. E. X-linked spermine synthase gene (SMS) defect: the first polyamine deficiency syndrome. *Eur. J. Hum. Genet.*, 11(12):937–44, 2003. ISSN 1018-4813. doi: 10.1038/sj.ejhg.5201072.

Cassiday, L. Structural biology: More than a crystallographer. *Nature*, 505(7485):711–713, 2014. ISSN 0028-0836. doi: 10.1038/nj7485-711a.

Catalá-López, F., Gènova-Maleras, R., Vieta, E., and Tabarés-Seisdedos, R. The increasing burden of mental and neurological disorders. *Eur. Neuropsychopharmacol.*, 2013. ISSN 1873-7862. doi: 10.1016/j.euroneuro.2013.04.001.

Catalá-López, F., Crespo-Facorro, B., Vieta, E., Valderas, J. M., Valencia, A., and Tabarés-Seisdedos, R. Alzheimer's disease and cancer: current epidemiological evidence for a mutual protection. *Neuroepidemiology*, 42(2):121–2, 2014a. ISSN 1423-0208. doi: 10.1159/000355899.

Catalá-López, F., Suárez-Pinilla, M., Suárez-Pinilla, P., Valderas, J. M., Gómez-Beneyto, M., Martinez, S., Balanzá-Martínez, V., Climent, J., Valencia, A., McGrath, J., Crespo-Facorro, B., Sanchez-Moreno, J., Vieta, E., and Tabarés-Seisdedos, R. Inverse and

direct cancer comorbidity in people with central nervous system disorders: a meta-analysis of cancer incidence in 577,013 participants of 50 observational studies. *Psychother. Psychosom.*, 83(2):89–105, 2014b. ISSN 1423-0348. doi: 10.1159/000356498.

Catts, V. S., Catts, S. V., O'Toole, B. I., and Frost, A. D. J. Cancer incidence in patients with schizophrenia and their first-degree relatives - a meta-analysis. *Acta Psychiatr. Scand.*, 117(5):323–36, 2008. ISSN 1600-0447. doi: 10.1111/j.1600-0447.2008.01163.x.

Cerny, V. Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm I. *J. Optim. Theory Appl.*, 45(l):41–51, 1985.

Chalmers, I. Randomized controlled trials of fetal monitoring 1973– 1977. In: Thalhammer O, Baumgarten K, Pollak A, eds. *Perinat. Med.*, pages 260–5, 1979.

Chalmers, T., Matta, R., Smith, H., and Kunzler, A.-M. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *NEJM*, 297: 1091–6, 1977.

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, 14:368, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-368.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, 41(Database issue):D816–23, 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1158.

Chen, W., Li, M., Wu, X., and Wang, J. Identifying protein complexes based on the integration of PPI network and gene expression data. *Int. J. Bioinform. Res. Appl.*, 11(1):30–44, 2015. ISSN 1744-5485. doi: 10.1504/IJBRA.2015.067337.

Chikina, M. D. and Sealfon, S. C. Increasing consistency of disease biomarker prediction across datasets. *PLoS One*, 9(4):e91272, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0091272.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19 Suppl 1:i84–90, 2003. ISSN 1367-4803.

Choura, M. and Rebaï, A. Topological features of cancer proteins in the human NR-RTK interaction network. *J. Recept. Signal Transduct. Res.*, 32(5):257–62, 2012. ISSN 1532-4281. doi: 10.3109/10799893.2012.702116.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007. ISSN 1744-4292. doi: 10.1038/msb4100180.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 1988.

Cohen, R., Erez, K., Ben-Avraham, D., and Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85(21):4626–8, 2000. ISSN 1079-7114.

Cooper, H. M. and Rosenthal, R. A comparison of statistical and traditional procedures for summarizing research. *Psychol Bull*, 87:442–449, 1980.

Cooper, H., Hedges, L., and Valentine, J. *The Handbook of Research Synthesis and Meta-Analysis*. 2009.

Cousin, E., Macé, S., Rocher, C., Dib, C., Muzard, G., Hannequin, D., Pradier, L., Deleuze, J.-F., Génin, E., Brice, A., and Campion, D. No replication of genetic association between candidate polymorphisms and Alzheimer's disease. *Neurobiol. Aging*, 32(8):1443–51, 2011. ISSN 1558-1497. doi: 10.1016/j.neurobiolaging.2009.09.004.

Crick, F. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970. ISSN 0028-0836.

Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–63, 1958. ISSN 0081-1386.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. Literature-curated protein interaction datasets. *Nat. Methods*, 6(1):39–46, 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1284.

D'Alche-Buc, F. and Wehenkel, L. Machine Learning in Systems Biology. *BMC Proc.*, 2 (Suppl 4):S1, 2008. ISSN 1753-6561.

Davies, D. L. and Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-1(2):224–227, 1979. ISSN 0162-8828. doi: 10.1109/TPAMI.1979.4766909.

Dayhoff, M. O. Computer analysis of protein evolution. *Sci. Am.*, 221(1):86–95, 1969. ISSN 0036-8733.

Dayhoff, M. O., Eck, R. V., Chang, M. A., and Sochard, M. R. *Atlas of Protein Sequence and Structure.* 1965.

Dayhoff, M. O. and Ledley, R. S. Comprotein. In *Proc. December 4-6, 1962, fall Jt. Comput. Conf. - AFIPS '62*, pages 262–274, New York, New York, USA, December 1962. ACM Press. doi: 10.1145/1461518.1461546. URL http://dl.acm.org/citation.cfm?id=1461518.1461546.

de Baumont, A., Maschietto, M., Lima, L., Carraro, D. M., Olivieri, E. H., Fiorini, A., Barreta, L. A. N., Palha, J. A., Belmonte-de Abreu, P., Moreira Filho, C. A., and Brentani, H. Innate immune response is differentially dysregulated between bipolar disease and schizophrenia. *Schizophr. Res.*, 161(2-3):215–21, 2015. ISSN 1573-2509. doi: 10.1016/j.schres.2014.10.055.

de Castro Leão, A., Duarte Dória Neto, A. a., and de Sousa, M. B. C. New developmental stages for common marmosets (Callithrix jacchus) using mass and age variables obtained by K-means algorithm and self-organizing maps (SOM). *Comput. Biol. Med.*, 39(10):853–9, 2009. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2009.05.009.

de Juan, D., Pazos, F., and Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, 14(4):249–261, 2013. ISSN 1471-0056. doi: 10.1038/nrg3414.

Devine, M. J., Plun-favreau, H., and Wood, N. W. Parkinson ' s disease and cancer : two wars , one front. 11(November):6386–6395, 2011.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15(2):330–40, 2005. ISSN 1088-9051. doi: 10.1101/gr.2821705.

Dobson, R. J. B., Munroe, P. B., Caulfield, M. J., and Saqi, M. A. S. Protein interaction networks associated with cardiovascular disease and cancer: exploring the effect of bias on shared network properties. *Int. J. Data Min. Bioinform.*, 9(4):339–57, 2014. ISSN 1748-5673.

Doolittle, R. F. and Blombaeck, B. Amino-acid sequence investigations of fibrinopeptides from various mammals: Evolutionary implications. *Nature*, 202:147–52, 1964. ISSN 0028-0836.

Doolittle, R. F., Schubert, D., and Schwartz, S. A. Amino acid sequence studies on artiodactyl fibrinopeptides. I. Dromedary camel, mule deer, and cape buffalo. *Arch. Biochem. Biophys.*, 118(2):456–67, 1967. ISSN 0003-9861.

Doolittle, R. F. The roots of bioinformatics in protein evolution. *PLoS Comput. Biol.*, 6 (7):e1000875, 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000875.

Doolittle, R. F., Oncley, J. L., and Surgenor, D. M. Species differences in the interaction of thrombin and fibrinogen. *J. Biol. Chem.*, 237:3123–7, 1962. ISSN 0021-9258.

Driver, J. a., Beiser, a., Au, R., Kreger, B. E., Splansky, G. L., Kurth, T., Kiel, D. P., Lu, K. P., Seshadri, S., and Wolf, P. a. Inverse association between cancer and Alzheimer's disease: results from the Framingham Heart Study. *Bmj*, 344(mar12 1):e1442–e1442, 2012. ISSN 0959-8138. doi: 10.1136/bmj.e1442.

Driver, J. A., Zhou, X. Z., and Lu, K. P. Pin1 dysregulation helps to explain the inverse association between cancer and Alzheimer's disease. *Biochim. Biophys. Acta*, 2015. ISSN 0006-3002. doi: 10.1016/j.bbagen.2014.12.025.

Duda, R., Hart, P., and Stork, D. *Pattern Classification*, volume 24. 2001. doi: 10.1007/s00357-007-0015-9.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. 1999.

Eddy, S. R. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998. ISSN 1367-4803. doi: 10.1093/bioinformatics/14.9.755.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 95(25):14863–8, 1998. ISSN 0027-8424.

El-Zayadi, A.-R. Hepatitis C comorbidities affecting the course and response to therapy. *World J. Gastroenterol.*, 15(40):4993–9, 2009. ISSN 2219-2840.

Eysenck, H. J. Systematic Reviews: Meta-analysis and its problems. *BMJ*, 309(6957): 789–792, 1994. ISSN 0959-8138. doi: 10.1136/bmj.309.6957.789.

Fan-Minogue, H., Chen, B., Sikora-Wohlfeld, W., Sirota, M., and Butte, A. J. A systematic assessment of linking gene expression with genetic variants for prioritizing candidate targets. *Pac. Symp. Biocomput.*, 20:383–94, 2015. ISSN 2335-6936.

Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19(2): 99–113, 1970. ISSN 0039-7989.

Fitch, W. M. The molecular evolution of cytochrome c in eukaryotes. *J. Mol. Evol.*, 8 (1):13–40, 1976. ISSN 0022-2844.

Fitch, W. M. and Margoliash, E. Construction of phylogenetic trees. *Science*, 155(3760): 279–84, 1967. ISSN 0036-8075.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269 (5223):496–512, 1995. ISSN 0036-8075.

Fortin, M., Bravo, G., Hudon, C., Vanasse, A., and Lapointe, L. Prevalence of multimorbidity among adults seen in family practice. *Ann. Fam. Med.*, 3(3):223–8, 2005. ISSN 1544-1717. doi: 10.1370/afm.272.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, R. D., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., and Venter, J. C. The minimal gene complement of Mycoplasma genitalium. *Science*, 270(5235):397–403, 1995. ISSN 0036-8075.

Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., Rogers, M., Ibanez, K., Boullosa, C., Andres Leon, E., Ben-Hur, A., and Valencia, A. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.*, 41(D1):D142–D151, 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1041.

Gal, G., Goral, A., Murad, H., Gross, R., Pugachova, I., Barchana, M., Kohn, R., and Levav, I. Cancer in parents of persons with schizophrenia: is there a genetic protection? *Schizophr. Res.*, 139(1-3):189–93, 2012. ISSN 1573-2509. doi: 10.1016/j.schres.2012.04.018.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–15, 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg405.

Geurts, P., Touleimat, N., Dutreix, M., and D'Alché-Buc, F. Inferring biological networks with output kernel trees. *BMC Bioinformatics*, 8 Suppl 2:S4, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S2-S4.

Glass, G. Primary, secondary and meta-analysis of research. *Educ Res.*, 10:3–8, 1976.

Glass, G., B, M., and ML, S. *Meta-Analysis in Social Research*. 1981.

Greengard, P., Valtorta, F., Czernik, A. J., and Benfenati, F. Synaptic vesicle phosphoproteins and regulation of synaptic function. *Science*, 259(5096):780–5, 1993. ISSN 0036-8075.

Hagen, J. B. The origins of bioinformatics. *Nat. Rev. Genet.*, 1(3):231–6, 2000. ISSN 1471-0056. doi: 10.1038/35042090.

Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)*. 2011.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. From molecular to modular cell biology. *Nature*, 1999.

Hasle, H., Clemmensen, I. H., and Mikkelsen, M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet*, 355(9199):165–9, 2000. ISSN 0140-6736. doi: 10.1016/S0140-6736(99)05264-2.

Haykin, S. *Neural Networks: A Comprehensive Foundation*. 1994.

Hedges, L. V. and Olkin, I. *Statistical Methods for Meta-analysis*. 1985.

Hernández-Avila, M., Lazcano-Ponce, E. C., Berumen-Campos, J., Cruz-Valdéz, A., Alonso de Ruíz, P. P., and Gonźalez-Lira, G. Human papilloma virus 16-18 infection

and cervical cancer in Mexico: a case-control study. *Arch. Med. Res.*, 28(2):265–71, 1997. ISSN 0188-4409.

Holloway, D. T., Kon, M., and DeLisi, C. Integrating genomic data to predict transcription factor binding. *Genome Inform.*, 16(1):83–94, 2005. ISSN 0919-9454.

Hong, F. and Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3): 374–82, 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm620.

Hosaka, M., Hammer, R. E., and Südhof, T. C. A phospho-switch controls the dynamic association of synapsins with synaptic vesicles. *Neuron*, 24(2):377–87, 1999. ISSN 0896-6273.

Huber, M. and Poulin, R. Antiproliferative effect of spermine depletion by N-cyclohexyl-1,3-diaminopropane in human breast cancer cells. *Cancer Res.*, 55 (4):934–43, 1995. ISSN 0008-5472.

Hudson, N. J., Reverter, A., and Dalrymple, B. P. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.*, 5(5):e1000382, 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000382.

Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., and Valencia, A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLoS Genet.*, 10(2): e1004173, 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004173.

Ibáñez, K., Guijarro, M., Pajares, G., and Valencia, A. A computational approach inspired by simulated annealing to study the stability of protein interaction networks in cancer and neurological disorders. *Data Min. Knowl. Discov.*, 2015. ISSN 1384-5810. doi: 10.1007/s10618-015-0410-5.

Ideker, T. and Sharan, R. Protein networks in disease. *Genome Res.*, 18(4):644–52, 2008. ISSN 1088-9051. doi: 10.1101/gr.071852.107.

Iyer, S., Killingback, T., Sundaram, B., and Wang, Z. Attack robustness and centrality of complex networks. *PLoS One*, 8(4):e59613, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0059613.

Jahchan, N. S., Dudley, J. T., Mazur, P. K., Flores, N., Yang, D., Palmerton, A., Zmoos, A.-F., Vaka, D., Tran, K. Q. T., Zhou, M., Krasinska, K., Riess, J. W., Neal, J. W., Khatri, P., Park, K. S., Butte, A. J., and Sage, J. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.*, 3(12):1364–77, 2013. ISSN 2159-8290. doi: 10.1158/2159-8290.CD-13-0183.

Jain, A. and Chandrasekaran, B. *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2. 1982. doi: 10.1016/S0169-7161(82)02042-2.

Jensen, L. J. and Bateman, A. The rise and fall of supervised machine learning techniques. *Bioinformatics*, 27(24):3331–2, 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr585.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000. ISSN 0028-0836. doi: 10.1038/35036627.

Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, 2001. ISSN 0028-0836. doi: 10.1038/35075138.

Ji, J., Sundquist, K., Ning, Y., Kendler, K. S., Sundquist, J., and Chen, X. Incidence of cancer in patients with schizophrenia and their first-degree relatives: a population-based study in Sweden. *Schizophr. Bull.*, 39(3):527–36, 2013. ISSN 1745-1701. doi: 10.1093/schbul/sbs065.

John E Hunter, F. L. S. and Jackson, G. B. *Meta-Analysis: Cumulating Research Findings Across Studies*. 1982.

Jonsson, P. F. and Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–7, 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl390.

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, 6(6):451–64, 2005. ISSN 1471-0056. doi: 10.1038/nrg1615.

Kamieniak, M. M., Rico, D., Milne, R. L., Muñoz Repeto, I., Ibáñez, K., Grillo, M. A., Domingo, S., Borrego, S., Cazorla, A., García-Bueno, J. M., Hernando, S., García-Donas, J., Hernández-Agudo, E., Y Cajal, T. R., Robles-Díaz, L., Márquez-Rodas, I., Cusidó, M., Sáez, R., Lacambra-Calvet, C., Osorio, A., Urioste, M., Cigudosa, J. C., Paz-Ares, L., Palacios, J., Benítez, J., and García, M. J. Deletion at 6q24.2-26 predicts longer survival of high-grade serous epithelial ovarian cancer patients. *Mol. Oncol.*, 2014. ISSN 1878-0261. doi: 10.1016/j.molonc.2014.09.010.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36(Database issue):D480–4, 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm882.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.*, 40(2):e15, 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1071.

Kaufman, A. C., Salazar, S. V., Haas, L. T., Yang, J., Kostylev, M. A., Jeng, A. T., Robinson, S. A., Gunther, E. C., van Dyck, C. H., Nygaard, H. B., and Strittmatter, S. M. Fyn inhibition rescues established memory and synapse loss in Alzheimer mice. *Ann. Neurol.*, 2015. ISSN 1531-8249. doi: 10.1002/ana.24394.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue):D841–6, 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1088.

Kim, T.-H., Choi, S. J., Lee, Y. H., Song, G. G., and Ji, J. D. Gene expression profile predicting the response to anti-TNF treatment in patients with rheumatoid arthritis; analysis of GEO datasets. *Joint. Bone. Spine*, 81(4):325–30, 2014. ISSN 1778-7254. doi: 10.1016/j.jbspin.2014.01.013.

Komurov, K. and Ram, P. T. Patterns of human gene expression variance show strong associations with signaling network hierarchy. *BMC Syst. Biol.*, 4:154, 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-154.

Komurov, K., White, M. A., and Ram, P. T. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput. Biol.*, 6(8), 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000889.

Laakso, M. and Hautaniemi, S. Integrative platform to translate gene sets to networks. *Bioinformatics*, 26(14):1802–3, 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq277.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I. n., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., and Robles, V. Machine learning in bioinformatics. *Brief. Bioinform.*, 7(1):86–112, 2006. ISSN 1467-5463.

Legány, C., Juhász, S., and Babos, A. Cluster validity measurement techniques. pages 388–393, 2006.

Li, J. and Tseng, G. C. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. 2011.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, 40(Database issue):D857–61, 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr930.

Lin, C.-Y., Lane, H.-Y., Chen, T.-T., Wu, Y.-H., Wu, C.-Y., and Wu, V. Y. Inverse association between cancer risks and age in schizophrenic patients: a 12-year nationwide cohort study. *Cancer Sci.*, 104(3):383–90, 2013a. ISSN 1349-7006. doi: 10.1111/cas.12094.

Lin, G.-M., Chen, Y.-J., Kuo, D.-J., Jaiteh, L. E. S., Wu, Y.-C., Lo, T.-S., and Li, Y.-H. Cancer incidence in patients with schizophrenia or bipolar disorder: a nationwide population-based study in Taiwan, 1997-2009. *Schizophr. Bull.*, 39(2):407–16, 2013b. ISSN 1745-1701. doi: 10.1093/schbul/sbr162.

Liou, Y.-C., Zhou, X. Z., and Lu, K. P. Prolyl isomerase Pin1 as a molecular switch to determine the fate of phosphoproteins. *Trends Biochem. Sci.*, 36(10):501–14, 2011. ISSN 0968-0004. doi: 10.1016/j.tibs.2011.07.001.

Littell, R. and Folks, J. Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Stat. Assoc.*, pages 802–806, 1971.

Liu, C., Che, D., Liu, X., and Song, Y. Applications of machine learning in genomics and systems biology. *Comput. Math. Methods Med.*, 2013:587492, 2013a. ISSN 1748-6718. doi: 10.1155/2013/587492.

Liu, T., Ren, D., Zhu, X., Yin, Z., Jin, G., Zhao, Z., Robinson, D., Li, X., Wong, K., Cui, K., Zhao, H., and Wong, S. T. C. Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform. *Sci. Rep.*, 3:3467, 2013b. ISSN 2045-2322. doi: 10.1038/srep03467.

Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat. Rev. Cancer*, 11(6):450–7, 2011. ISSN 1474-1768. doi: 10.1038/nrc3063.

Lu, K. P. Pinning down cell signaling, cancer and Alzheimer's disease. *Trends Biochem. Sci.*, 29(4):200–9, 2004. ISSN 0968-0004. doi: 10.1016/j.tibs.2004.02.002.

Ma, L.-L., Yu, J.-T., Wang, H.-F., Meng, X.-F., Tan, C.-C., Wang, C., and Tan, L. Association between cancer and Alzheimer's disease: systematic review and meta-analysis. *J. Alzheimers. Dis.*, 42(2):565–73, 2014. ISSN 1875-8908. doi: 10.3233/JAD-140168.

Manczinger, M. and Kemény, L. Novel factors in the pathogenesis of psoriasis and potential drug candidates are found with systems biology approach. *PLoS One*, 8 (11):e80751, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0080751.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, I. H. W. The WEKA Data Mining Software: An Update. *SIGKDD Explor.*, 11(1), 2009.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37 (Database issue):D619–22, 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn863.

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., and Irizarry, R. A. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, 39(Database issue):D1011–5, 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1259.

McCall, M. N., Jaffee, H. a., and Irizarry, R. a. fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. *Bioinformatics*, 28(23):3153–4, 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts588.

Milanesi, L., Romano, P., Castellani, G., Remondini, D., and Liò, P. Trends in modeling Biomedical Complex Systems. *BMC Bioinformatics*, 10 Suppl 1:I1, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S12-I1.

Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. Towards a detailed atlas of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 23(6):929–40, 2013. ISSN 1879-033X. doi: 10.1016/j.sbi.2013.07.005.

Murga, M. and Fernández-Capetillo, O. Genomic instability: on the birth and death of cancer. *Clin. Transl. Oncol.*, 9(4):216–20, 2007. ISSN 1699-048X.

Musicco, M., Adorni, F., Di Santo, S., Prinelli, F., Pettenati, C., Caltagirone, C., Palmer, K., and Russo, A. Inverse occurrence of cancer and Alzheimer disease: A population-based incidence study. *Neurology*, 81(4):322–8, 2013. ISSN 1526-632X. doi: 10.1212/WNL.0b013e31829c5ec1.

Negrini, S., Gorgoulis, V. G., and Halazonetis, T. D. Genomic instability–an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.*, 11(3):220–8, 2010. ISSN 1471-0080. doi: 10.1038/nrm2858.

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J. M., and Pascual-Montano, A. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, 37(Web Server issue):W317–22, 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp416.

Ochoa, D., Juan, D., Valencia, A., and Pazos, F. Detection of significant protein coevolution. *Bioinformatics*, 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv102.

Okerlund, N. D. and Cheyette, B. N. R. Synaptic Wnt signaling-a contributor to major psychiatric disorders? *J. Neurodev. Disord.*, 3(2):162–74, 2011. ISSN 1866-1955. doi: 10.1007/s11689-011-9083-6.

Olteanu, M. and Villa-Vialaneix, N. On-line relational and multiple relational som. *Neurocomputing*, 147:15–30, 2015.

Olteanu, M., Villa-Vialaneix, N., and Cottrell, M. On-line relational som for dissimilarity data. In *Advances in Self-Organizing Maps (Proceedings of WSOM 2012, Santiago, Chili, 12-14 decembre 2012), Estevez P., Principe J., Zegers P., Barreto G. (eds.), Advances in Intelligent Systems and Computing series*, volume 198, pages 13–22, Berlin/Heidelberg, 2012. Springer Verlag.

Omer, A., Singh, P., Yadav, N. K., and Singh, R. K. An overview of data mining algorithms in drug induced toxicity prediction. *Mini Rev. Med. Chem.*, 14(4):345–54, 2014. ISSN 1875-5607.

O'Rourke, K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *J. R. Soc. Med.*, 100(12):579–82, 2007. ISSN 0141-0768. doi: 10.1258/jrsm.100.12.579.

Ou, S.-M., Lee, Y.-J., Hu, Y.-W., Liu, C.-J., Chen, T.-J., Fuh, J.-L., and Wang, S.-J. Does Alzheimer's disease protect against cancers? A nationwide population-based study. *Neuroepidemiology*, 40(1):42–9, 2013. ISSN 1423-0208. doi: 10.1159/000341411.

Pajares, G. and de la Cruz, J. On combining support vector machines and simulated annealing in stereovision matching. *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, 34 (4):1646–57, 2004. ISSN 1083-4419.

Pearson, H. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006. ISSN 1476-4687. doi: 10.1038/441398a.

Pearson, K. Report on Certain Enteric Fever Inoculation Statistics. *Br. Med. J.*, 2(2288): 1243–6, 1904. ISSN 0007-1447.

Pennisi, E. Genomics. DNA study forces rethink of what it means to be a gene. *Science*, 316(5831):1556–7, 2007. ISSN 1095-9203. doi: 10.1126/science.316.5831.1556.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U. S. A.*, 96(16):9212–7, 1999. ISSN 0027-8424.

Pons, T., Paramonov, I., Boullosa, C., Ibáñez, K., Rojas, A. M., and Valencia, A. A common structural scaffold in CTD phosphatases that supports distinct catalytic

mechanisms. *Proteins*, 82(1):103–18, 2014. ISSN 1097-0134. doi: 10.1002/prot.24376.

Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, 39(11):1338–49, 2007. ISSN 1546-1718. doi: 10.1038/ng.2007.2.

Quinlan, J. R. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986. ISSN 0885-6125. doi: 10.1007/BF00116251.

Quinlan, J. R. C4.5: programs for machine learning. 1993.

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.*, 28 (7):323–32, 2012. ISSN 0168-9525. doi: 10.1016/j.tig.2012.03.004.

Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A., and Ciccarelli, F. D. Low duplicability and network fragility of cancer genes. *Trends Genet.*, 24(9):427–30, 2008. ISSN 0168-9525. doi: 10.1016/j.tig.2008.06.003.

Raspopovic, J., Marcon, L., Russo, L., and Sharpe, J. Modeling digits. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science*, 345(6196):566–70, 2014. ISSN 1095-9203. doi: 10.1126/science.1252960.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. Meta-Analysis of Microarrays : Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer Meta-Analysis of Microarrays : Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Ca. *Cancer Res.*, pages 4427–4433, 2002.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S., Yang, X., Ghamsari, L., Balcha, D., Begg, B., Braun, P., Brehme, M., Broly, M., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A.,

Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B., Hardy, M., Jin, M., Kang, S., Kiros, R., Lin, G., Luck, K., MacWilliams, A., Menche, J., Murray, R., Palagi, A., Poulin, M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J., Scholz, A., Shah, A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A., Trigg, S., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M., Xia, Y., Barabási, A.-L., Iakoucheva, L., Aloy, P., DeLasRivas, J., Tavernier, J., Calderwood, M., Hill, D., Hao, T., Roth, F., and Vidal, M. A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159(5):1212–1226, 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.10.050.

Romero, J. P., Benito-León, J., Louis, E. D., and Bermejo-Pareja, F. Alzheimer's disease is associated with decreased risk of cancer-specific mortality: a prospective study (NEDICES). *J. Alzheimers. Dis.*, 40(2):465–73, 2014. ISSN 1875-8908. doi: 10.3233/JAD-132048.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408, 1958. ISSN 0033-295X.

Rosenthal, R. Combining results of independent studies. *Psychol. Bull.*, 85(1):185–193, 1978.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24(3):227–35, 2000. ISSN 1061-4036. doi: 10.1038/73432.

Rost, B. and Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, 1994. ISSN 0887-3585. doi: 10.1002/prot.340190108.

Rozenberg, G., Bck, T., and Kok, J. N. Handbook of Natural Computing. 2011.

Rung, J. and Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, 14(2):89–99, 2013. ISSN 1471-0064. doi: 10.1038/nrg3394.

S. Kirkpatrick, C. D. G., Vecchi, M. P., Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science (80-.).*, 220(4598):671–80, 1983. ISSN 0036-8075. doi: 10.1126/science.220.4598.671.

Sachlos, E., Risueño, R. M., Laronde, S., Shapovalova, Z., Lee, J.-H., Russell, J., Malig, M., McNicol, J. D., Fiebig-Comyn, A., Graham, M., Levadoux-Martin, M., Lee, J. B., Giacomelli, A. O., Hassell, J. a., Fischer-Russell, D., Trus, M. R., Foley, R., Leber, B., Xenocostas, A., Brown, E. D., Collins, T. J., and Bhatia, M. Identification of drugs including a dopamine receptor antagonist that selectively target cancer stem cells. *Cell*, 149(6):1284–97, 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.03.049.

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., and Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. from bed to bench*, 7(1):17–31, 2014. ISSN 2008-2258.

Saigo, H., Vert, J., Akutsu, T., and Ueda, N. Comparison of SVM-Based Methods for Remote Homology Detection. *Genome Informatics*, page 396, 2011. doi: 10.11234/gi1990.13.396.

Sánchez-Lladó, F. J., Pajares, G., and López-Martínez, C. Improving the Wishart Synthetic Aperture Radar image classifications through Deterministic Simulated Annealing. *ISPRS J. Photogramm. Remote Sens.*, 66(6):845–857, 2011. ISSN 09242716. doi: 10.1016/j.isprsjprs.2011.09.007.

Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–23, 2009. ISSN 1476-4687. doi: 10.1038/nature08454.

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PLoS One*, 7(2):e31826, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0031826.

Schramm, G., Kannabiran, N., and König, R. Regulation patterns in signaling networks of cancer. *BMC Syst. Biol.*, 4:162, 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-162.

Shamir, R. and Sharan, R. Algorithmic Approaches to Clustering Gene Expression Data. *Curr. Top. Comput. Mol. Biol.*, MIT Press:269–300, 2002.

Sheth, R., Marcon, L., Bastida, M. F., Junco, M., Quintana, L., Dahn, R., Kmita, M., Sharpe, J., and Ros, M. A. Hox genes regulate digit patterning by controlling the wavelength of a Turing-type mechanism. *Science*, 338(6113):1476–80, 2012. ISSN 1095-9203. doi: 10.1126/science.1226804.

Shi, F., Abraham, G., Leckie, C., Haviv, I., and Kowalczyk, A. Meta-analysis of gene expression microarrays with missing replicates. *BMC Bioinformatics*, 12(1):84, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-84.

Solé, R. V., Valverde, S., Rodriguez-Caso, C., and Sardanyés, J. Can a minimal replicating construct be identified as the embodiment of cancer? *Bioessays*, 36(5):503–12, 2014. ISSN 1521-1878. doi: 10.1002/bies.201300098.

Song, C. and Tseng, G. C. HYPOTHESIS SETTING AND ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS. *Ann. Appl. Stat.*, 8(2):777–800, 2014. ISSN 1932-6157.

Srihari, S., Madhamshettiwar, P. B., Song, S., Liu, C., Simpson, P. T., Khanna, K. K., and Ragan, M. A. Complex-based analysis of dysregulated cellular processes in cancer. *BMC Syst. Biol.*, 8 Suppl 4:S1, 2014. ISSN 1752-0509. doi: 10.1186/1752-0509-8-S4-S1.

Stjernswärd, J. Decreased survival related to irradiation postoperatively in early breast cancer. *Lancet*, 304:1285–6, 1974.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, 10(9):2997–3011, 1982. ISSN 0305-1048. doi: 10.1093/nar/10.9.2997.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–50, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102.

Sun, J. and Zhao, Z. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genomics*, 11 Suppl 3:S5, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-S3-S5.

Tabarés-Seisdedos, R. and Rubenstein, J. L. R. Chromosome 8p as a potential hub for developmental neuropsychiatric disorders: implications for schizophrenia, autism and cancer. *Mol. Psychiatry*, 14(6):563–89, 2009. ISSN 1476-5578. doi: 10.1038/mp.2009.2.

Tabarés-Seisdedos, R. and Rubenstein, J. L. Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders. *Nat. Rev. Neurosci.*, 14(April): 293–304, 2013. ISSN 1471-0048. doi: 10.1038/nrn3464.

Tabarés-seisdedos, R. and Rubenstein, J. L. Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders. *Nat. Rev. Neurosci.*, 14(April): 293–304, 2013.

Tabarés-Seisdedos, R., Dumont, N., Baudot, A., Valderas, J. M., Climent, J., Valencia, A., Crespo-Facorro, B., Vieta, E., Gómez-Beneyto, M., Martínez, S., and Rubenstein, J. L. No paradox, no progress: inverse cancer comorbidity in people with other complex diseases. *Lancet Oncol.*, 12(6):604–8, 2011. ISSN 1474-5488. doi: 10.1016/S1470-2045(11)70041-9.

Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res.*, 40(Web Server issue):W478–83, 2012. ISSN 1362-4962. doi: 10.1093/nar/gks402.

Taminau, J., Lazar, C., Meganck, S., and Nowé, A. Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis. *ISRN Bioinforma.*, 2014:1–7, 2014. ISSN 2090-7346. doi: 10.1155/2014/345106.

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., and Drăghici, S. Machine learning and its applications to biology. *PLoS Comput. Biol.*, 3(6):e116, 2007. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030116.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, 27(2):199–204, 2009. ISSN 1546-1696. doi: 10.1038/nbt.1522.

Teschendorff, A. E. and Severini, S. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst. Biol.*, 4(1):104, 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-104.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22 (19):2405–12, 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl406.

Thampi, S. M. Introduction to bioinformatics. *CoRR*, abs/0911.4230, 2009. URL http://arxiv.org/abs/0911.4230.

Thinnes, F. P. Why cancer survivors have a lower risk of Alzheimer disease. *Mol. Genet. Metab.*, 107(3):630–1, 2012. ISSN 1096-7206. doi: 10.1016/j.ymgme.2012.06.016.

Thompson, J. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, 25(24): 4876–4882, 1997. ISSN 13624962. doi: 10.1093/nar/25.24.4876.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994. ISSN 0305-1048. doi: 10.1093/nar/22.22.4673.

Tseng, G. C., Ghosh, D., and Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, 40(9):3785–99, 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1265.

Turing, A. M. The Chemical Basis of Morphogenesis. *Philos. Trans. R. Soc. B Biol. Sci.*, 237(641):37–72, 1952. ISSN 0962-8436. doi: 10.1098/rstb.1952.0012.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. Defining comorbidity: implications for understanding health and health services. *Ann. Fam. Med.*, 7(4):357–63, 2009. ISSN 1544-1717. doi: 10.1370/afm.983.

Van Noorden, R., Maher, B., and Nuzzo, R. The top 100 papers. *Nature*, 514(7524): 550–3, 2014. ISSN 1476-4687. doi: 10.1038/514550a.

van Pel, D. M., Barrett, I. J., Shimizu, Y., Sajesh, B. V., Guppy, B. J., Pfeifer, T., McManus, K. J., and Hieter, P. An Evolutionarily Conserved Synthetic Lethal Interaction Network Identifies FEN1 as a Broad-Spectrum Target for Anticancer Therapeutic Development. *PLoS Genet.*, 9(1):e1003254, 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003254.

Vanitha, C. D. A., Devaraj, D., and Venkatesulu, M. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Comput. Sci.*, 47:13–21, 2015. ISSN 18770509. doi: 10.1016/j.procs.2015.03.178.

Wachi, S., Yoneda, K., and Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–8, 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti688.

Wadhwa, N., Mathew, B. B., Jatawa, S. K., and Tiwari, A. Genetic instability in urinary bladder cancer: An evolving hallmark. *J. Postgrad. Med.*, 59(4):284–8, 2013. ISSN 0022-3859. doi: 10.4103/0022-3859.123156.

Wang, X., Kang, D. D., Shen, K., Song, C., Lu, S., Chang, L.-C., Liao, S. G., Huo, Z., Tang, S., Ding, Y., Kaminski, N., Sibille, E., Lin, Y., Li, J., and Tseng, G. C. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, 28(19):2534–6, 2012. ISSN 1367-4811.

Wang, Z., Chen, Y., and Li, Y. A brief review of computational gene prediction methods. *Genomics. Proteomics Bioinformatics*, 2(4):216–21, 2004. ISSN 1672-0229.

Warnat, P., Eils, R., and Brors, B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-265.

West, J., Bianconi, G., Severini, S., Teschendorff, A. E., and Genomics, S. C. Differential network entropy reveals cancer system hallmarks. *Sci. Rep.*, 2:802, 2012. ISSN 2045-2322. doi: 10.1038/srep00802.

Won, K.-J., Hamelryck, T., Prügel-Bennett, A., and Krogh, A. An evolutionary method for learning HMM structure: prediction of protein secondary structure. *BMC Bioinformatics*, 8:357, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-357.

Wu, C. H. and McLarty, J. W. Neural Networks and Genome Informatics. 2000.

Wu, C. H. Artificial neural networks for molecular sequence analysis. *Comput. Chem.*, 21(4):237–256, 1997. ISSN 00978485. doi: 10.1016/S0097-8485(96)00038-1.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. Integrated network analysis platform for protein-protein interactions. *Nat. Methods*, 6(1):75–77, 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1282.

Wuchty, S. and Almaas, E. Peeling the yeast protein network. *Proteomics*, 5(2):444–9, 2005. ISSN 1615-9853. doi: 10.1002/pmic.200400962.

Xia, J., Sun, J., Jia, P., and Zhao, Z. Do cancer proteins really interact strongly in the human protein-protein interaction network? *Comput. Biol. Chem.*, 35(3):121–5, 2011. ISSN 1476-928X. doi: 10.1016/j.compbiolchem.2011.04.005.

Xiao, Q., Wang, J., Peng, X., Wu, F.-X., and Pan, Y. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. *BMC Genomics*, 16 Suppl 3:S1, 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S3-S1.

Xiong, W., Xie, L., Zhou, S., Liu, H., and Guan, J. The centrality of cancer proteins in human protein-protein interaction network: a revisit. *Int. J. Comput. Biol. Drug Des.*, 7(2-3):146–56, 2014. ISSN 1756-0756. doi: 10.1504/IJCBDD.2014.061643.

Xu, L., Geman, D., and Winslow, R. L. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, 8:275, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-275.

Xu, L., Tan, A. C., Winslow, R. L., and Geman, D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics*, 9:125, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-125.

Yang, Z. R. Biological applications of support vector machines. *Brief. Bioinform.*, 5(4): 328–338, 2004. ISSN 1467-5463. doi: 10.1093/bib/5.4.328.

Yao, P. J., Zhu, M., Pyun, E. I., Brooks, A. I., Therianos, S., Meyers, V. E., and Coleman, P. D. Defects in expression of genes related to synaptic vesicle trafficking in frontal cortex of Alzheimer's disease. *Neurobiol. Dis.*, 12(2):97–109, 2003. ISSN 0969-9961.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.4.309.

Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics*, 10(6):402–15, 2009. ISSN 1875-5488. doi: 10.2174/138920209789177575.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., and Liebler, D. C. Proteogenomic characterization of human

colon and rectal cancer. *Nature*, 513(7518):382–7, 2014. ISSN 1476-4687. doi: 10.1038/nature13438.

Zilliox, M. J. and Irizarry, R. A. A gene expression bar code for microarray data. *Nat. Methods*, 4(11):911–3, 2007. ISSN 1548-7091. doi: 10.1038/nmeth1102.